Proceedings of the Indian Academy of Sciences

Vol. 89, January-December 1980

CONTENTS (Mathematical Sciences)

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 1–24.

Criteria for the unitarizability of some highest weight modules*

R PARTHASARATHY
School of Mathematics, Tata Institute of Fundamental Research, Bombay 400 005

MS received 20 April 1979

Abstract. For a linear semisimple Lie group we obtain a necessary and sufficient condition for a highest weight module with non-singular infinitesimal character to be unitarizable.

Keywords. Highest weight module; spin module; formal Dirac operator; parabolic subalgebra; noncompact roots; infinitesimal character.

1. Introduction

A linear semisimple Lie group G admits nontrivial unitarizable highest weight modules precisely when it admits holomorphic discrete series. Supposing G is such a group, it is of interest to characterise the unitarizable ones among the set of all highest weight modules of G. We are looking for a condition which is both necessary and sufficient for a highest weight module π of G to be unitarizable. The most desirable (and one that would be the simplest) is to give a condition directly on the highest weight μ of the module π. The main results of this paper (theorem A, § 3 and theorem B, § 5) give such an explicit necessary and sufficient condition on μ, provided the infinitesimal character of π is nonsingular. In § 6, we discuss the applications of our results to the (o, p) Betti numbers of compact quotients of bounded domains.

Let G be a connected linear semisimple Lie group and let G_c be the complexification of G. Assume that G_c is simply connected. Let g_0 be the Lie algebra of G and let g be the Lie algebra of G_c. Let K be a maximal compact subgroup of G. Let k_0 be the Lie algebra of K and k the complexification of k_0. Let $g_0 = k_0 + p_0$ be the Cartan decomposition and let $g = k + p$ be its complexification. We will denote by θ the corresponding Cartan involution.

We now assume that the symmetric space G/K is a hermitian symmetric domain. As is well-known p can be canonically identified with the space of (complex) tangent vectors at the identity coset eK in G/K. Let p_+ be the subspace of p consisting of the holomorphic tangent vectors at eK and p_- the space of antiholomorphic tangent vectors at eK. It is well known that both p_+ and p_- are K submodules of p. Let b be a Cartan subgroup of k and r_k a Borel subalgebra of k containing b. Then one knows that b is a Cartan subalgebra of g and that $r_k + p_+$ is a Borel subalgebra of g.

* Partially supported by a grant from NSF (MCS 76-05962).

Let r be this Borel subalgebra of g. Let $\triangle$ be the set of roots of g with respect to b and let $\triangle_k$ and $\triangle_n$ be the sets of compact and non-compact roots respectively, so that

$$k = b + \Sigma_{\alpha \in \triangle_k} g^\alpha \text{ and}$$

$$p = \Sigma_{\alpha \in \triangle_n} g^\alpha$$

here g^α denotes the one dimensional space spanned by a root vector corresponding to α. Let P be the set of positive roots defined by the Borel subalgebra r. Thus

$$(1.1) \qquad r = b + \Sigma_{\alpha \in P} g^\alpha.$$

P_k and P_n will denote respectively the set of compact and non-compact roots in P. We denote by δ, δ_k and δ_n half the sum of the roots in P, P_k and P_n respectively.

Let $U(g)$ be the enveloping algebra of g and $U(k)$ be the enveloping algebra of k. Let π be an irreducible smooth representation of G in a space $\tilde{H}$. Let H be the space of K finite vectors in $\tilde{H}$ so that $U(g)$ has an irreducible representation π in H for which H is $U(k)$ finite.

(1.2) *Definition* : π is said to be a highest weight module for g (or for G) if there exists $\mu \in b^*$ and $v \in H$, $v \neq 0$ such that

(1) For every $T \in b$, $\pi(T)v = \mu(T)v$,

(2) $\pi(X_\alpha)v = 0$ if $\alpha \in P_k$ or $\alpha \in -P_n$.

When there is some confusion, we will specify π is a highest weight module with respect to $P_k \cup -P_n$.

(1.3) *Definition* : π is said to be unitarizable if there exists a positive definite inner product $(,)$ on H, such that for every X in g_0,

$$(\pi(X)v, w) + (v, \pi(X)w) = 0 \text{ for all } v, w \text{ in } H.$$

Problem : Describe the set of all highest weight modules for G which are unitarizable.

We will focus our attention on the set of all highest weight modules for G which have a nonsingular infinitesimal character (see § 3 for definition).

If π is a highest weight module for G, then upto equivalence π is uniquely determined by its highest weight μ (Definition (1·2)) and μ is uniquely determined by π. Also μ satisfies

$$(1.4) \qquad 2(\mu, \alpha)/(\alpha, \alpha) \in \mathbf{Z} \text{ for every } \alpha \in P$$

and $$2(\mu, \alpha)/(\alpha, \alpha) \in \mathbf{Z}^+ \text{ for every } \alpha \in P_k.$$

Moreover, to every μ satisfying (1.4) there corresponds (upto equivalence) a unique highest weight module π_μ of G whose highest weight is μ. This module is obtained as follows : Let V_μ be the finite dimensional irreducible module for K with highest weight μ. Regard V_μ as a module for $k + p_-$ by making the action of p_- trivial. Then the highest weight module π_μ is simply the unique irreducible quotient of $U(g) \otimes_{U(k+p_-)} V_\mu$.

We denote by H_μ this irreducible quotient. One can show that the action π_μ on H_μ comes as the action on K finite vectors of a suitable irreducible representation of G.

We note that $P_k \cup -P_n$ is the set of positive roots with respect to another lexicographic ordering on $\triangle$. We denote by $\bar{P}$ this system of positive roots. Recall the g Verma modules $V_{g,\bar{P},\eta}$ with highest weight η (an element of b) relative to $\bar{P}$. Let $Z(g)$ denote the centre of the enveloping algebra $U(g)$. As is well-known elements of $Z(g)$ act by scalar multiplication on $V_{g,\bar{P},\eta}$. Let $\chi_{\bar{P},\eta}$ denote the corresponding homomorphism of $Z(g)$ into C. If ω denotes the Casimir element in $Z(g)$, then it is known that

$$\chi_{\bar{P},\eta}(\omega) = (\eta + \delta_{\bar{P}},\ \eta + \delta_{\bar{P}}) - (\delta_{\bar{P}},\ \delta_{\bar{P}})$$

where $\delta_{\bar{P}}$ denotes half the sum of the roots in $\bar{P}$. Thus

(1.5) $$\chi_{\bar{P},\eta}(\omega) = (\eta - \delta_n + \delta_k,\ \eta - \delta_n + \delta_k) - (\delta,\ \delta)$$

as $(\delta,\ \delta) = (\bar{\delta},\ \bar{\delta})$

(1.6) *Corollary* : The Casimir acts on the highest weight module H_μ by the scalar

$$(\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k) - (\delta,\ \delta)$$

Proof : It is not hard to see that H_μ is precisely the irreducible quotient of $V_{g,\bar{P},\mu}$. Hence the result (q.e.d.).

2. An inequality satisfied for unitarizable representations

Let $\pi = \pi_\mu$ be an irreducible highest weight module for G. Let $H = H_\mu$ be the space of K finite vectors. *We assume henceforth that π is unitarizable.* For the results to be stated in this section, π can be an arbitrary irreducible *unitary* representation of G. Let L, L^+ and L^- be the spin module and the two half spin modules for $so(p)$ the Lie algebra of the special orthogonal group $SO(p)$ (The symmetric bilinear form on $p \subset g$ is the restriction of the Killing form). By composing with the adjoint action of k on p, we obtain the spin representation σ of k on L and the two half-spin representations $\sigma^\pm$ of k on L^+ and L^-. Recall that for every $x \in p$ there is a clifford multiplication $c(x) : L \to L$. Now, $H \otimes L$ is a k module and we have a formal Dirac operator $D : H \otimes L \to H \otimes L$ defined by

(2.1) $$D = \Sigma\, \pi(X_i) \otimes c(X_i).$$

Here the summation is over an orthonormal basis for p_0. There is a unique (upto a positive scalar multiple) positive definite inner product $(,)$ on L such that for every X in p_0 and s, s' in L,

(2.2) $$(c(x)\, s,\ s') + (s,\ c(x)\, s') = 0.$$

Since we have a positive definite inner product on H, for which also

(2.3) $$(\pi(x)\, u,\ v) + (u,\ \pi(x)\, v) = 0$$

for every x in g_0 and for u, v in H, we now have a positive definite inner product on $H \otimes L$, the product of the ones on H and L. For u, v in H and s, s' in L, then

$$(u \otimes s,\ v \otimes s') = (u,\ v)\ (s,\ s').$$

With respect to this inner product we clearly have

(2.4) $(Dw, w') = (w, Dw')$

for w, w' in $H \otimes L$.

Let ω_k be the Casimir element in $U(k)$. It equals $-\Sigma Y_i^2$ where Y_i is a basis of k_0 such that $(Y_i, Y_j) = \delta_{ij}$ where $(,)$ denotes the Killing form of g_0.

A formula was obtained in [3, § 3] for the square of the Dirac operator. These computations also apply to the square of the formal Dirac operator (of also [7, § 8]). One thus obtains

(2.5) *Lemma* : $D^2 = (\pi \otimes \sigma)(\omega_k) - \pi(\omega) \otimes 1 - (\delta, \delta) + (\delta_k, \delta_k)$.

(2.6) *Proposition*: Assume that ξ is the highest weight of an irreducible k submodule of $H \otimes L$.

Then

$$(\xi + \delta_k,\ \xi + \delta_k) \geqslant (\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k).$$

Proof : Let w be an element of $H \otimes L$ contained in an irreducible k submodule of $H \otimes L$ with highest weight ξ. The Casimir ω_k of k acts on w by the scalar $(\xi + \delta_k,\ \xi + \delta_k) - (\delta_k, \delta_k)$. Thus by (2.5) and (1.6)

$$\begin{aligned} D^2 w &= (\xi + \delta_k,\ \xi + \delta_k) - (\delta_k, \delta_k) - (\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k) \\ &\quad + (\delta, \delta) - (\delta, \delta) + (\delta_k, \delta_k) \\ &= (\xi + \delta_k,\ \xi + \delta_k) - (\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k). \end{aligned}$$

Hence

(2.7) $(D^2 w,\ w) = \{(\xi + \delta_k,\ \xi + \delta_k) - (\mu - \delta_n + \delta_k)\}\ (w, w)$.

But $(D^2 w, w) = (DDw, w) = (Dw, Dw)$.

The last quantity is non-negative since the hermitian form on $H \otimes L$ is positive definite. For the same reason (w, w) is also nonnegative. Hence from (2.7) the assertion in the proposition follows.

(2.8) *Corollary* : Let π_μ be an irreducible highest weight module for G. Assume π_μ is unitarizable. Let V_μ be the irreducible finite dimensional module of k with highest weight μ. Suppose ξ is the highest weight of an irreducible k submodule of $V_\mu \otimes L$. Then,

$$(\xi + \delta_k, \xi + \delta_k) \geqslant (\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k).$$

Proof : This is clear from (2.6) since $V_\mu \subseteq H_\mu$.

3. A condition on μ

Let (π_μ, H_μ) be a highest weight module for G. In § 1, we observed that the centre $Z(g)$ of $U(g)$ acts on the Verma module $V_{\sigma, \bar{p}, \eta}$ by the homomorphism $\chi_{\bar{p}, \eta} : Z(g) \to C$. Any homomorphism χ of $Z(g)$ into C is of the form $\chi_{\bar{p}, \eta}$ for a suitable element η in b^*. The homomorphism χ is said to be nonsingular if $\eta + \delta_{\bar{p}}$ is nonsingular, i.e. $(\eta + \delta_{\bar{p}}, \alpha) \neq 0$ for any root α.

We will now assume that the infinitesimal character of π_μ is nonsingular. Since the infinitesimal character of π_μ is given by the homomorphism $\chi_{\tilde{P},\mu}$, our assumption amounts to making the hypothesis $\mu - \delta_n + \delta_k$ is nonsingular, i.e.

(3.1) $(\mu - \delta_n + \delta_k, \alpha) \neq 0$ for any root α.

In addition, recall that the highest weights μ of highest weight modules for G satisfy the condition

$2(\mu, \alpha)/(\alpha, \alpha) \in \mathbf{Z}$ for every α in P and

$2(\mu, \alpha)/(\alpha, \alpha) \in \mathbf{Z}^+$ for every α in P_k.

Let P' be the set of roots defined by

(3.2) $P' = [\alpha \in \triangle \mid (\mu - \delta_n + \delta_k, \alpha) > 0]$.

Note that P' is the set of positive roots with respect to a lexicographic ordering. Also, observe that

(3.3) $2(\mu - \delta_n + \delta_k, \alpha)/(\alpha, \alpha)$ is a positive integer for every α in P'.

Let $\delta' =$ half the sum of the roots in P'. Then, note that

(3.4) $\mu - \delta_n + \delta_k = \lambda + \delta'$.

Where λ satisfies

(3.5) $2(\lambda, \alpha)/(\alpha, \alpha)$ is a non-negative integer for every α in P'.

For every α in P_k, $(\mu, \alpha) \geqslant 0$, $(-\delta_n, \alpha) = 0$ and $(\delta_k, \alpha) > 0$.

Hence

(3.6) $P' \supseteq P_k$.

Let P'_n be the set of non-compact roots in P' and let δ' and let δ'_n be half the sum of the roots in P'_n.

Then $\delta' = \delta_k + \delta'_n$ and so (3.4) implies

(3.7) $\mu = \lambda + \delta_n + \delta'_n$

Using our assumption that π_μ is unitarizable, we wish to conclude that the quantities λ and P'_n appearing above have very special properties. We now introduce some more terminology to explain this.

(3.8) Recall that r was the Borel subalgebra of g corresponding to the positive system P. Let q be a parabolic subalgebra of g containing r. Let

$$q = m + u$$

be the Levi decomposition of q such that m contains b. Thus u is the unipotent radical of q and m is a reductive component of q. Let P_m be the roots of (m, b) which are contained in P. Let P_n be the roots of P whose corresponding root-

spaces are contained in u. Thus P is the disjoint union of P_m and P_u. If P^- denotes the set $(-P_m) \cup P_u$ then it is known that P^- is also a positive system. Set P_n^- to be the set of non-compact roots in P^-.

(3.9) From the assumption that π_μ is unitarizable, we wish to conclude the following property about the expression $\mu = \lambda + \delta_n + \delta_n'$. *There exists a parabolic-subalgebra* q *containing* r *such that with the notation introduced above* $P_n' = P_n^-$ *and* $(\lambda, a) = 0$ *for every* a *in* P_m.

(3.10) *Example* : Observe that if $(\mu - \delta_n + \delta_k, a) > 0$ for every a in P, then $P = P'$ and the above property is easily seen to hold by taking $q = r$ the Borel subalgebra itself. (P_m = empty in this case and $P_n^- = P_n$). This is precisely the case when π_μ is a member of the holomorphic discrete series. As another illustration, we mention the case $\mu = 0$, so that π_μ is the trivial one dimensional representation, which is unitarizable. In this case, the property is seen to hold by taking $q = g$ and $\lambda = 0$ ($P_m = P$ in this case and $P_n^- = -P_n$).

(3.11) *We will quickly see that* $(\lambda, a) = 0$ *for every* a *in* $P \cap -P'$. Corollary (2·8) says that if ξ is the highest weight of an irreducible k submodule of $V_\mu \otimes L$, then $(\xi + \delta_k, \xi + \delta_k) \geqslant (\mu - \delta_n + \delta_k, \mu - \delta_n + \delta_k)$. The spin module L is self dual and one has knowledge about the highest weights of irreducible k submodules of L (cf. [3, §2]). Using these one can see that the irreducible k module with lowest weight $-\delta_n'$ occurs in L. Let us denote by $V_{-\delta_n'}$ this component contained in L. Then $V_\mu \otimes V_{-\delta_n'} \subseteq V_\mu \otimes L$. If we now take $\xi = \mu - \delta_n$, then since $\mu = \lambda + \delta_n + \delta_n'$ (cf. (3.7)), $\mu - \delta_n' \subseteq = \lambda + \delta_n$. Both λ and δ_n are dominant and integral with respect to P_k. Thus there is an irreducible finite dimensional k module $V_{\mu - \delta_n'}$ with highest weight $\mu - \delta_n'$. By [6, 2·26] $V_{\mu - \delta_n'}$ occurs in $V_\mu \otimes V_{-\delta_n'}$. Applying corollary (2.8) to $\xi = \mu - \delta_n' = \lambda + \delta_n$ we conclude that

$$(\lambda + \delta_n + \delta_k, \lambda + \delta_n + \delta_k) \geqslant (\mu - \delta_n + \delta_k, \mu - \delta_n + \delta_k)$$

Since $\mu = \lambda + \delta_n + \delta_n'$, $\mu - \delta_n + \delta_k = \lambda + \delta_n' + \delta_k = \lambda + \delta'$ (cf. (3.4)).

Thus

$$(\lambda + \delta_n + \delta_k, \lambda + \delta_n + \delta_k) \geqslant (\lambda + \delta', \lambda + \delta')$$

i.e. $$(\lambda + \delta, \lambda + \delta) \geqslant (\lambda + \delta', \lambda + \delta')$$

i.e. $$(\lambda, \lambda) + 2(\lambda, \delta) + (\delta, \delta) \geqslant (\lambda, \lambda) + 2(\lambda, \delta') + (\delta', \delta').$$

But $(\delta, \delta) = (\delta', \delta')$. So we conclude that

$$(\lambda, \delta) \geqslant (\lambda, \delta')$$

That is $$(\lambda, \delta' - \delta) \leqslant 0.$$

But λ is dominant with respect to P'. Hence we conclude that $(\lambda, a) = 0$ for every a in $P \cap -P'$.

We wanted to show that there exists a parabolic subalgebra q containing r such that $P_n' = P_n^-$ and $(\lambda, a) = 0$ for every a in P_m. For any q if we *define*

$$P_{m,n} = \text{the set of non-compact roots in } P_m$$

then

$$P_{m,n} = P_n \cap -P_n^-.$$

Also, the assertion that $P_n' = P_n^-$ is the same as the assertion

(3.12) $\quad P_{m,n} = P_n \cap -P_n'.$

We will now commence a long chain of arguments and eventually show that μ is of the very special type discussed in (3.9).

Suppose Y is a subset of $P_n \cap -P_n'$. *We denote by* q_Y the intersection of all parabolic subalgebras q of g containing r such that P_m contains Y.

(3.13) *Remark.* For each $Y \subseteq P_n \cap -P_n'$, q_Y has the following property. No semisimple ideal of the reductive part of q_Y is contained in k.

The reason is the following. As is well-known the parabolic subalgebras q of g containing r are in one to one correspondence with subsets of the set S of simple roots of P. Suppose q contains r and suppose $Y \subseteq P_m$, where P_m is the set of roots of P which belong to the reductive part m of q. Suppose m has a semisimple ideal m_1 such that $m_1 \subseteq k$. This is equivalent to the statement " Let $X \subseteq S$ be the subset of S corresponding to q. Then X can be written as a disjoint union $X_1 \cup X_2$ such that all the roots of X_1 are compact and X_1 is orthogonal to X_2, i.e. $(\alpha, \beta) = 0$ for any α in X_1 and any β in X_2." But then if q_2 is the parabolic subalgebra of g containing r corresponding to $X_2 \subseteq S$, then q_2 is a proper subalgebra of q and the set of non-compact roots in the reductive part of q_2 is exactly the same as those in the reductive part of q. In particular, Y is still contained in the set of roots of the reductive part of q_2, since Y contains only non-compact roots. Since q_Y is the intersection of all parabolic subalgebras q of g containing r for which $Y \subseteq P_m$, it is now clear that the reductive part of q_Y has no semisimple ideal contained in k. This completes the proof of (3.13).

Varying Y over the subsets of $P_n \cap -P_n'$ we get a collection of parabolic subalgebras

$$\{q_Y \mid Y \subseteq P_n \cap -P_n'\}.$$

It should be remarked that the set $P_{m,n}$ of those non-compact roots of P which are contained in the reductive part of any $q = q_Y$ *may not* be contained in $P_n \cap -P_n'$.

(3.14) Consider the collection of those parabolic subgroups $q = q_Y$, $Y \subseteq P_n \cap -P_n'$. for which $P_{m,n}$ *is contained* in $P_n \cap -P_n'$. (This set is non-empty since q_Y, when Y is the empty set is obviously a member). *Among all such q_Y, choose one for which $P_{m,n}$ has maximum possible cardinality.*

In what follows, unless otherwise stated, q will denote this *particular* parabolic subalgebra and the sets $P_{m,n}$, P_m, P_n^-, etc. (cf. (3.8)) shall all be with respect to this particular parabolic subalgebra q. We now set

(3.15) $\quad P^1 = P_k \cup P_n^-$

(3.16) We claim that P^1 is a positive system in $\triangle$.

The simplest way to prove this is through the following argument. Under the well-known one to one correspondence between parabolic subalgebras of g contain-

ing r and subsets of S, there is a unique subset X of S corresponding to our parabolic subalgebra q chosen. (The subset X is precisely the set of simple roots of P_m). Suppose for a while g is not only semisimple but actually simple. (This assumption is not really necessary but is made here only to illustrate the argument. The proof in the general case is alike). Then one knows that S contains only one simple non-compact root, say α_1. Also it is known that the coefficient of α_1 in the highest root in P is one. This is *characteristic* of the hermition symmetric case, where $P_k \cup P_n$ and also $P_k \cup -P_n$ are *both* positive systems. If α_1 belongs to X (which, as was observed before, is the set of simple roots for $P_m = P_{m,k} \cup P_{m,n}$) then α_1 is the only non-compact root in X and its coefficient in the highest root of P_m is one. Thus $P_{m,k} \cup (-P_{m,n})$ is a positive system for the set of roots of m with respect to b. If α_1 does not belong to X then $P_{m,n}$ is empty and so again $P_{m,k} \cup (-P_{m,n})$ is a positive system for the roots of m. If $\tilde{P}_m$ is any positive system for the roots of $\tilde{m}$, then $P_m \cup P_u$ (where P_u is the set of all roots in P, whose root-spaces are contained in the unipotent radical of q) is a positive system for the roots of b in g. But one sees easily that the set P^1 in (3.15) is nothing but

(3.17) $P^1 = (P_{m,k} \cup -P_{m,n}) \cup P_u$ (disjoint).

Thus, the assertion (3.16) is proved. As we remarked, the case when g is not simple can be treated in the same way.

(3.18) We now let S^1 be the set of simple roots of P^1. Let r^1 be the Borel subalgebra of g corresponding to P^1. Since $P^1 = P_k \cup P_n^-$ (the latter defined with respect to q) one sees at once that r^1 is contained in q.

(3.19) Let X^1 be the subset of S^1 corresponding to q.

(3.20) We enumerate S^1 as $\alpha_1, \alpha_2, \cdots, \alpha_i, \cdots, \alpha_j$ in such a way that $X^1 = \alpha_1, \alpha_2, \cdots \alpha_i$.

Remark : Even when g is simple S^1 may contain more than one non-compact root.

We now show that $P' = P^1$. This will be used in the proof of 3.9.

(3.21) Suppose P' is not equal to P^1.

We wish to show that (3.21) leads to a contradiction, namely (3.34).

If P^1 is not equal to P', then there is a simple root α in S^1, such that $-\alpha$ belongs to P'.

However P^1 and P' have some common parts. Let us look at this very carefully. First of all both P^1 and P' contain P_k (cf. (3·15) and (3·6)). Secondly, observe that by the choice of q made in (3·14), $P_n \cap P_n' \subseteq P_{u,n}$, where $P_{u,n}$ denotes the set of non-compact roots, whose root spaces are contained in the unipotent radical of q. But $P_{u,n} \subseteq P^1$ (cf. (3·17)). So, $P_n \cap P_n' \subseteq P_n^1$, where P_n^1 denotes the set of non-compact roots in P^1. This means that all those roots which are common to $-P_n$ and $-P_n'$ are also common to $-P_n'$ and $-P_n^1$. In particular, a root common to $-P_n'$ and P_n^1 cannot be in $-P_n \cap -P_n'$; it has to lie in

$P_n \cap -P_n'$. The root a picked out in the beginning of this paragraph is not common to P^1 and P'. Thus, in view of the preceding observations, we can infer two facts about this root a. First, it cannot be a compact root; thus it has to lie in $P_n^1 \cap -P_n'$. So, secondly it cannot be in $-P_n \cap -P_n'$ but has to be in $P_n \cap -P_n'$. Without loss, we can assume that S^1 has been enumerated in (3.20), in such a way that $a = a_{i+1}$.

Thus,

(3.22) *In the enumeration* (3.20), a_{i+1} is a *noncompact root and* $a_{i+1} \in P_n \cap -P_n'$.

For any positive integer e such that $1 \leqslant e \leqslant j$, let q^e denote the parabolic subalgebra of g corresponding to the subset $\{a_1, a_2, \cdots, a_e\}$ of S^1. Thus $q^1 = q$ (cf. (3·20)) and q^{i+1} denotes the parabolic subalgebra of g corresponding to the subset $\{a_1, a_2, \cdots, a_{i+1}\}$ of S^1. Note that since q^{i+1} contains $q^i = q$, *afortiori.*

(3.23) q^{i+1} contains the Borel subalgebra r.

(3.24) *We also claim that* q^{i+1} *is of the form* q_Y *described before* (3.13) *for a suitable subset* $Y \subseteq P_n \cap -P_n'$.

In fact let Y^0 be the set $P_{m,n}$ of the non-compact roots in P which belong to the reductive part of our chosen q. Then $Y^0 \subseteq P_n \cap -P_n'$ and $q = q_{Y^0}$. If we now let $Y = Y^0 \cup \{a_{i+1}\}$ then in view of (3.22) $Y \subseteq P_n \cap -P_n'$ and it is easy to see that q^{i+1} equals the corresponding q_Y. Let P_m^{i+1} denote the set of roots of the reductive part of q^{i+1} which belong to P. *Let* $P_{m,n}^{i+1}$ *denote the set of non-compact roots in* P_m^{i+1}. The reductive part of q^{i+1} contains the reductive part of $q^i = q$ and is strictly bigger than the reductive part of q ; a_{i+1} is a nonc-ompact root which belongs to the reductive part of q^{i+1} but it does not belong to the reductive part of q^i. Thus, in fact, the set of non-compact roots in the reductive part of q^i is a proper subset of the set of non-compact roots in the reductive part of q^{i+1}. But q was chosen to be maximal having a certain property stated in (3.14). Thus we conclude

(3.25) $P_{m,n}^{i+1}$ *is not contained in* $P_n \cap -P_n'$.

But, evidently, by very definition, $P_{m,n}^{i+1} \subseteq P_n$. Thus we conclude that

(3.26) *there is a root* β *of* $P_{m,n}^{i+1}$ *which belongs to* $P_n \cap P_n'$.

The root β will be the 'trump' in our '*reductio ad absurdum.*' Since $P_n \cap P_n' \subseteq P^1$, (cf. arguments preceding (3.22)) β is a root in P^1, hence a non-negative integral linear combination of the simple roots S^1 of P^1. In particular,

(3.27) $\beta = A + da_{i+1}$, where A is a non-negative integral linear combination of the roots $a_1, \cdots, a_i$ and d is a positive integer.

Note that many of the roots in $\{a_1, \ldots, a_i\}$ may be compact. To proceed with the argument, we would like to show that A *can actually be written as a nonnegative real linear combination of the set of non-compact roots of a positive system for* $\triangle_m$, the roots of the reductive part m of q. To this end we will prove a slightly more general result.

(3.28) *Lemma.* Suppose g_0 is any real semisimple Lie algebra. Let $g_0 = k_0 + p_0$ be a Cartan decomposition of g_0. Assume that g_0 has no semisimple ideals contained in k_0. Let b_0 be a Cartan subalgebra of k_0 and assume b_0 is also a Cartan subalgebra of g_0. Let g, k, b, etc. be the complexifications. Let $\triangle$ be the set of roots of (g, b). Let ϕ be any real linear form on $i\,b_0$. Then there exists a positive system P in $\triangle$ such that ϕ is a non-negative real linear combination of elements of P_n, the set of non-compact roots in P.

Proof. Start with *any* positive system P^0 in $\triangle$. Let S^0 be the set of simple roots of P^0. Let $S_0 = A_1 \cup A_2 \cup \cdots \cup A_t$ be a partition of S^0 such that

A_t = all the non-compact roots in S^0,

A_{t-1} = all those compact roots in S^0, which are connected (i.e. having a non-zero scalar product) with some element of A_t,

A_{t-2} = all those compact roots in $S^0 - \{A_t \cup A_{t-1}\}$ which are connected to A_{t-1}

A_{t-3} = all those compact roots in $S^0 - A_t \cup A_{t-1} \cup A_{t-2}$ which are connected to A_{t-2}, etc.

Because of the assumption that g_0 has no compact factors the above procedure certainly exhausts all of S^0.

Let $\alpha_1, \alpha_2, \alpha_3, \ldots$ be an enumeration of elements of A_1, $\beta_1, \beta_2, \beta_3, \ldots$ an enumeration of elements of A_2, etc.

Let $\phi = \Sigma_{\gamma \in S^0} m_\gamma \gamma$ be the unique expression for ϕ in terms of the basis elements $\{\gamma \mid \gamma \in S^0\}$; m_γ are real numbers, some negative and some non-negative. Without loss of generality we can assume that m_{α_1}, the coefficient of α_1 in ϕ is non-negative. Let q^1 be the parabolic subalgebra of g corresponding to the subset $S^0 - \{\alpha_1\}$ of S^0. Let g^1 be the reductive part of q^1 and u^1, the unipotent radical of q^1. Observe that *there is at least one non-compact root of* P^0 *occurring in* u^1; for, otherwise, $p \subseteq g^1$ $(\neq g)$ which can only happen if g_0 has compact semisimple ideals, contrary to what was assumed. Let ξ be a noncompact root of P^0 occurring in u^1. In particular, ξ can be written as a non-negative integral linear combination of elements of S^0, such that the coefficient of α_1 is positive. We now choose a non-negative real number c such that if

(3.29) $\phi - c\xi = \Sigma_{\gamma \in S^0} n_\gamma \gamma$,

then $n_{\alpha_1} = 0$.

(3.30) The complex Lie algebra g^1 is the complexification of the real Lie algebra $g_0^1 = g^1 \cap g_0$ and g_0^1 has no compact semisimple factors.

Let us postpone the proof of this but assume it for a while.

By (3.29), $\phi - c\,\xi$ is a real linear combination of roots of g^1. Also, the semisimple rank of g^1 is strictly less than the semisimple rank of g. Thus, using a suitable induction hypothesis, we can assume that $\phi - c\xi$ is a non-negative real linear combination of the non-compact roots of some positive system $P^{(1)}$ of the roots of

g^1 with respect to b. This can be done because by (3·30), g_0^1 (of which g^1 is the complexification) satisfies the hypothesis of the lemma.

All we have to do now to prove the lemma is to enlarge $P^{(1)}$ to a positive system P of the roots of g with respect to b, by adjoining the roots which occur in the unipotent radical of q^1. For then (3.29) achieves our aim.

It now remains to show (3.30) is true.

Suppose g_c is a compact form of a complex semisimple Lie algebra g and suppose q is a parabolic subalgebra of g. Then the intersection of q with g_c is a compact real form of a reductive part of q. This is well known. In our case $g_c = k_0 + iP_0$ is a compact real form of g. Let k_0^1 and p_0^1 be the intersection of g^1 with k_0 and p_0 respectively. We observe that q^1 and g^1 are both stable under the Cartan involution θ associated to the Cartan decomposition $g_0 = k_0 + p_0$. (The reason is this: since we assumed rank of g_0 = rank of k_0, θ is the inner automorphism of an element of $\exp b_0$; but $b_0 \subseteq m^1 \subseteq q^1$). In view of this remark it is not hard to see that the intersection of g^1 with g_0 is $k_0^1 + P_0^1$ and the intersection of g^1 with $g_c\,(= k_0 + ip_0)$ is $k_0^1 + ip_0^1$. Thus $g^1 \cap g_0$ is a real form of g^1 since $g^1 \cap g_c$ is so.

Now the real reductive Lie algebra $g_0^1 = k_0^1 + p_0^1$ has a Cartan subalgebra b_0 contained in k_0^1 and we can talk of compact and non-compact roots. The set $S^0 - \{a_1\}$ is the set of simple roots for an appropriate positive system for the roots of g^1 with respect to b. If g^1 has a semisimple ideal contained in k^1, then $S^0 - \{a_1\}$ can be written as a disjoint union $X_1 \cup X_2$ such that all the roots of X_1 are compact and X_1 is orthogonal to X_2. But this cannot be done as is seen by the way the a_1 was chosen.

This completes the proof of lemma (3.28).

Applying the result of (3.28) to the quantity A on the right hand side of the equality (3.27), we see that there is a positive system Q for the roots of m (the reductive part of the parabolic subalgebra q chosen in (3.14)) such that

$$A = \Sigma_{\gamma \in Q_n} m_\gamma \gamma^\gamma$$

where γ runs through the set Q_n of non-compact roots in Q and m_γ are non-negative real numbers. Thus from (3.27) we obtain

$$(3.31) \qquad \beta = d a_{i+1} + \Sigma_{\gamma \in Q_n} m_\gamma \gamma^\gamma.$$

Where d, m_γ are all non-negative real numbers.

By the choice of q (cf. (3·14)), for every non-compact root α in the reductive part of q, either α or $-\alpha$ lies in $P_n \cap -P_n'$. Thus,

(3.32) If $Q_n = \{\gamma_1, \gamma_2, \ldots, \gamma_r\}$ then either γ_i or $-\gamma_i$ lies in $P_n \cap -P_n'$.

We now enlarge Q to a positive system Q^* for the roots of g, by adjoining to Q the set P_u of roots in the unipotent radical of q. It should be remarked that Q^* may not contain P_k. Let δ_k^* (resp. δ_n^*) be half the sum of the compact roots (resp. non-compact roots) in Q^*. There is a unique element w of the Weyl group of k such

$$(3.33) \qquad \delta_k = w^{-1} \delta_k^*.$$

Then $w^{-1}\,\delta_n^*$ is the highest weight of an irreducible component of the spin module L for k. Since L is selfdual, $w^{-1}\,(-\delta_n^*)$ is the lowest weight of an irreducible component of the spin module L. Clearly $-\delta_n^*$ is in the orbit under the Weyl group of k of this lowest weight. We denote this irreducible component by $V_{-\delta_n^*}$. The contradiction we are aiming at (from 3.21) is the following :

(3.34) There is an irreducible component V_ξ with highest weight ξ contained in $V_{\lambda+\delta_n+\delta_n'} \otimes V_{-\delta_n^*} \subseteq V_{\lambda+\delta_n+\delta_n'} \otimes L$ for which $(\xi+\delta_k,\ \xi+\delta_k)$ is strictly less than $(\lambda+\delta',\ \lambda+\delta')$ (compare with Corollary (2.8) and (3.4)).

The proof of (3·34) is also a little lengthy. We proceed as follows.

Let V_ϕ and V_τ be two irreducible finite dimensional modules for k with highest weights (with respect to P_k) ϕ and τ respectively. Let s_1 and s_2 be two elements of the Weyl group W_k of k. For any root a let s_a denote the element of W_k which corresponds to the reflection associated to a. Suppose there is an element $a \in P_k$ such that

$$s_2 = s_a s_1$$

and $N(s_2) = N(s_1) + 1$, where $N(s)$ denotes the length of s, i.e., the length of a minimal expression for s as the product of reflections associated to simple roots in P_k. For any $s \in W_k$, let $V_{\phi+s\tau}$ denote the unique irreducible finite dimensional representation whose highest weight lies in the orbit of $\phi + s\tau$. Let ω_k be the Casimir element in the enveloping algebra of k. Let C_s denote the constant by which ω_k acts on $V_{\phi+s\tau}$. We claim

(3.35) $\quad C_{s_2} \leqslant C_{s_1}$.

Also, let t be the unique element of the Weyl Group W_k such that $t(P_k) = -P_k$. Then we also claim

(3.36) $C_t \leqslant C_s$ for any $s \in W_k$.

To show (3.35), it is enough to show that $\phi + s_2\tau$ is a weight of the irreducible. module $V_{\phi+s_1\tau}$. The latter fact will follow from the following computation. On the one hand,

$$(3.37) \quad s_a(\phi + s_1\tau) = s\,a\,\phi + s_2\,\tau$$
$$= \phi + s_2\tau - \frac{2(\phi, a)}{(a, a)}\,a.$$

On the other hand, $s_a(\phi + s_1\,\tau) = \phi + s_1\tau - 2\,[(\phi, a)/(a, a)]\,a - 2\,[(s_1\tau, a)/a, a)]\,a$. Therefore, using (3.37)

$$\phi + s_2\,\tau = \phi + s_1\tau - 2\,\frac{(s_1\tau, a)}{(a, a)}\,a$$

Since τ is dominant with respect to P_k and since $N(s_a\,s_1) > N(s_1)$, $2\,(s_1\tau, a)/(a, a)$ is a nonnegative integer. Both $\phi + s_1\tau$ and $s_a(\phi + s_1\tau)$ are weights of $V_{\phi+s_1\tau}$. From what we said above and from (3.37) it follows that $\phi + s_2\tau$ is in between $s_a(\phi + s_1\tau)$ and $\phi + s_1\tau$ in the a-string of weights of $V_{\phi+s_1\tau}$ through $\phi + s_1\tau$. But the a-string of weights through a given weight of an irreducible module is

unbroken. Therefore $\phi + s_2\tau$ is a weight of $V_{\phi+s_1\tau}$. Thus, the claim (3.35) is proved.

Applying (3.35) successively (3·36) follows.

We will now apply (3.35) and (3.36) to $\phi = \lambda + \delta_n + \delta'_n$ and $\tau =$ the unique element in the orbit of $-\delta^*_n$ which is the highest weight of $V_{-\delta^*_n}$. In particular we can conclude the following.

(3.38) Let $s \in W_k$ be such that $s^{-1}(-\delta^*_n)$ is the highest weight of $V_{-\delta^*_n}$. Let $t \in W_k$ be such that $tP_k = -P_k$. Let C_s (resp. C_t) be the value of Casimir ω_k on $V_{\lambda+\delta_n+\delta'_n-\delta^*_n}$, the representation of k whose highest weight belongs to the orbit of $\lambda + \delta_n + \delta'_n - \delta^*_n$ (resp. value of ω_k on $V_{\lambda+\delta_n+\delta'_n+ts^{-1}(-\delta^*_n)}$).
Then

(3.39) $C_t \leqq C_s$.

The *crucial observation in concluding the proof of* (3.34), is the following lemma:

(3.40) *Lemma.* Let s be defined as in (3.38) and let C_s be the value of the Casimir ω_k on $V_{\lambda+\delta_n+\delta'_n-\delta^*_n}$. Then $C_s + (\delta_k, \delta_k)$ is strictly less than $(\mu - \delta_n + \delta_k, \mu - \delta_n + \delta_k)$.

(Recall $\mu = \lambda + \delta_n + \delta'_n$).

Proof: The root β chosen in (3.26) will play the key role in the proof. To understand the argument, we first consider the case $\lambda = 0$ and we investigate the value C_s of the Casimir ω_k on $V_{\delta_n+\delta'_n-\delta^*_n}$.

Let $\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_r$ be the roots in P_n and let $-\beta_1, -\beta_2, \ldots, -\beta_j, \beta_{j+1}, \ldots, \beta_r$ be the roots in P'_n.
Recall the positive system Q for the roots of m, the reductive part of q, and the positive system Q^* for the roots of g, which was obtained by adjoining to Q the set P_u of roots in the unipotent radical of q. The set Q_n of noncompact roots in Q is described in (3.32). Also it is clear that every root in the unipotent radical of q belongs to P, since the Borel subalgebra of g defined by P is contained in q. From these descriptions, it is clear that the set Q^*_n, the set of non-compact roots in Q is contained in $\{\pm\beta_1, \pm\beta_2, \ldots, \pm\beta_j, \beta_{j+1}, \cdots, \beta_r\}$. The root α_{i+1} ((cf. 3.22)) is contained in $P_n \cap -P$. That is

$$\alpha_{i+1} \in \{\beta_1, \beta_2, \ldots, \beta_j\}.$$

Also, by our choice (cf. 3.20, 3.22) α_{i+1} is *not* a root in the set of non-compact roots in the reductive part m of q. Thus we can and do arrange the enumeration $\{\beta_1, \beta_2, \cdots, \beta_j, \cdots, \beta_r\}$ of P_n, so that in addition to the properties already mentioned, we also have,

(3.41) $Q_n = \{-\beta_1, -\beta_2, \ldots, -\beta_e, \beta_{e+1}, \ldots, \beta_f\}$

where $f < j$

(3.42) $Q^*_n = Q_n \cup \{\beta_{f+1}, \ldots, \beta_j, \beta_{j+1}, \ldots, \beta_r\}$ and $\beta_j = \alpha_{i+1}$ (cf. (3.22)).

In addition we observe that the root β chosen in (3.26) belongs to $P_n \cap P'_n = \{\beta_{j+1}, \ldots, \beta_r\}$. We can assume without loss that β is enumerated to be β_{j+1}.

Applying these notation and using (3.31) and (3.41) we obtain

$$(3.43)\quad \beta_{j+1} = -a_1\beta_1 - \ldots - a_e\beta_e + a_{e+1}\beta_{e+1} + a_f\beta_f + a_j\beta_j$$

where a_i are non-negative real numbers.

With these preparations, we can now get back to analysing the value of the Casimir ω_k on $V_{\delta_n+\delta'_n-\delta^*_n}$ the irreducible representation whose highest weight lies in the orbit of $\delta_n + \delta'_n - \delta^*_n$. Note that

$$\delta_n = \tfrac{1}{2}(\beta_1 + \beta_2 + \ldots + \beta_r)$$

$$\delta'_n = \tfrac{1}{2}(-\beta_1 - \ldots - \beta_j + \beta_{j+1} + \ldots + \beta_r)$$

$$\delta^*_n = \tfrac{1}{2}(-\beta_1 \ldots - \beta_e + \beta_{e+1} \ldots + \beta_r).$$

Hence $\delta_n + \delta'_n - \delta^*_n$ is given by

$$(3.44)\quad \delta_n + \delta'_n - \delta^*_n = \tfrac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} - \ldots - \beta_f - \ldots - \beta_j + \beta_{j+1} \ldots + \beta_r).$$

Several observations must be made from the expression on the right hand side of the equality in (3.44). *First of all it shows that* $\delta_n + \delta'_n - \delta^*_n$ *is a weight of the spin module* L *for* k (cf. [3, § 2]. Moreover,

$$(3.45)\quad \tfrac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} - \ldots - \beta_f - \ldots - \beta_j + \beta_{j+1} + \ldots + \beta_r)$$ is *not* in the orbits under W_k of the highest weights of irreducible components of L.

As we will see below, (3.45) will essentially follow from (3.43). A weight ϕ of the spin module L is in the orbit of the highest weight of some irreducible component of L if and only if

$$(3.46)\quad \phi = \tfrac{1}{2}(\gamma_1 + \ldots + \gamma_r)$$

where $\{\gamma_1, \ldots, \gamma_r\}$ is the set of noncompact roots in some positive system for the roots of g. It follows easily from (3.43) that whenever $\{\beta_1, \ldots, \beta_e, -\beta_{e+1}, \ldots, -\beta_f, -\beta_j\}$ is contained in a set $\{\gamma_1, \gamma_2, \ldots, \gamma_r\}$ as described above then $-\beta_{j+1}$ also belongs to $\{\gamma_1, \ldots, \gamma_r\}$. In particular, $\{\beta_1, \ldots, \beta_e, -\beta_{e+1}, \ldots, -\beta_f, \ldots, -\beta_j, \beta_{j+1}, \ldots, \beta_r\}$ cannot be the set of non-compact roots of a positive system for the roots of g. This is enough to conclude that $\tfrac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} \ldots - \beta_f \ldots - \beta_j + \beta_{j+1} + \ldots + \beta_r)$ is not in the W_k orbit of the highest weight of any irreducible component of L. One might wonder why can't $\tfrac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} \ldots - \beta_f \ldots - \beta_j + \beta_{j+1} \ldots + \beta_r)$ equal $\tfrac{1}{2}(\gamma_1 + \ldots + \gamma_r)$ where $\gamma_1, \ldots, \gamma_r$ is a set as described after (3.46). But if it were so, that would make the multiplicity of $\tfrac{1}{2}(\gamma_1 + \ldots + \gamma_r)$ as a weight of L equal to at least two (cf. [3, § 2]) which by [3, § 2] again cannot happen.

Thus (3.45) is proved and hence by (3.44) $\delta_n + \delta'_n - \delta^*_n$ is a weight of the spin module L for k, but $V_{\delta_n+\delta'_n-\delta^*_n}$ is *not* an irreducible component of L.

We state now a general fact. Suppose ϕ is a weight of an irreducible finite dimensional module V_τ with highest weight τ. Assume that ϕ is not in the orbit of τ. Let V_ϕ be the irreducible module whose highest weight lies in the orbit of ϕ. Then the value of the Casimir ω_k on V_ϕ is strictly less than the value of ω_k on V_τ. To see this let $s\phi$ be the highest weight of V_ϕ where s is an element of W_k. Then $s\phi + \Sigma m_\alpha \alpha = \tau$ where $\Sigma m_\alpha \alpha$ is a nonnegative integral linear combination of the

roots in P_k, with at least one m_α different from zero. Thus $(\tau + \delta_k, \tau + \delta_k) = (s\phi + \delta_k, s\phi + \delta_k) + 2(s\phi + \delta_k, \Sigma m_\alpha \alpha) + (\Sigma m_\alpha \alpha, \Sigma m_\alpha \alpha)$ which is strictly greater than $(s\phi + \delta_k, s\phi + \delta_k)$. Our assertion follows from this.

On every irreducible component of L, ω_k acts by $(\delta, \delta) - (\delta_k, \delta_k)$. Thus we conclude that,

(3.47) ω_k acts on $V_{\delta_n + \delta'_n - \delta^*_n}$ by a constant strictly less than $(\delta, \delta) - (\delta_k, \delta_k)$. This completes the proof of lemma (3.40) in the case $\lambda = 0$. In the general case we argue as follows.

Consider the irreducible finite dimensional module $V_{\lambda + \delta_n + \delta'_n - \delta^*_n}$ discussed preceeding (3.39).

$$\lambda + \delta_n + \delta'_n - \delta^*_n = \lambda + \tfrac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} \ldots - \beta_f \ldots$$
$$- \beta_j + \beta_{j+1} + \ldots + \beta_r)$$

Let F_λ be the finite dimensional irreducible module for g whose highest weight lies in the orbit (under the Weyl group of g) of λ. Consider the k module $F_\lambda \otimes L$. Let $\tilde{P}$ be a positive system for the roots of g. Let $\tilde{\lambda}$ be the highest weight of F_λ with respect to $\tilde{P}$ and let $\tilde{\delta}_n$ be half the sum of the non-compact roots in $\tilde{P}$. Let $V_{\tilde{\lambda} + \tilde{\delta}_n}$ be the irreducible module for k whose highest weight lies in the orbit (under W_k) of $\tilde{\lambda} + \tilde{\delta}_n$. For each $\tilde{P}$, $V_{\tilde{\lambda} + \tilde{\delta}_n}$ occurs in $F_\lambda \otimes L$. On each one of the modules $V_{\tilde{\lambda} + \tilde{\delta}_n}$, the Casimir ω_k acts by the same constant, namely, $(\tilde{\lambda} + \tilde{\delta}, \tilde{\lambda} + \tilde{\delta}) - (\tilde{\delta}_k, \tilde{\delta}_k)$. For any other irreducible component V_ξ of $F_\lambda \otimes L$, the action of ω_k on V_ξ is strictly less than the above constant. No element in the orbit of $\lambda + \frac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} \ldots - \beta_f - \ldots - \beta_j + \beta_{j+1} + \ldots + \beta_r)$ can be of the form $\tilde{\lambda} + \tilde{\delta}_n$ as described above, for the same reasons as we saw for the case $\lambda = 0$. Since $\lambda + \delta_n + \delta'_n - \delta^*_n$ is equal to $\lambda + \frac{1}{2}(\beta_1 + \ldots + \beta_e - \beta_{e+1} \ldots - \beta_f \ldots - \beta_j + \beta_{j+1} + \ldots + \beta_r)$ we conclude that ω_k acts on $V_{\lambda + \delta_n + \delta'_n - \delta^*_n}$ by a constant strictly less than $(\tilde{\lambda} + \tilde{\delta}, \tilde{\lambda} + \tilde{\delta}) - (\delta_k, \delta_k)$. The latter constant is simply $(\lambda + \delta', \lambda + \delta') - (\delta_k, \delta_k)$ since λ is the highest weight of F_λ with respect to P'. Thus the lemma (3.40) is completely proved.

Looking at (3.39) and the lines preceeding it and using lemma (3.40) we now conclude the following :

Let V_ξ be the irreducible component of $V_{\lambda + \delta_n + \delta'_n} \otimes V_{-\delta^*_n}$, whose highest weight ξ lies in the orbit of the sum of $\lambda + \delta_n + \delta'_n$ and $ts^{-1}(-\delta^*_n)$ which are respectively the highest weight of $V_{\lambda + \delta_n + \delta'_n}$ and the lowest weight of $V_{-\delta^*_n}$ (cf. [6]). Then the Casimir ω_k acts on V_ξ by a constant which is strictly less than $(\lambda + \delta', \lambda + \delta') - (\delta_k, \delta_k)$. In other words $(\xi + \delta_k, \xi + \delta_k)$ is strictly less than $(\lambda + \delta', \lambda + \delta')$.

This completes the proof of (3.34).

In view of corollary (2.8) the statement (3.34) clearly implies a contradiction. Thus, the assumption (3.21), namely, that P' is not equal to P^1 (c.f. (3.15) for the definition of P^1) leads to (3.34) which in turn leads to a contradiction. Thus, we have proved $P' = P^1$. In particular, $P'_n = P^1_n$, which by (3.15) is equal to P^-_n, the latter being defined with respect to q (cf. (3.8)).

We have thus obtained a very explicit necessary condition on the parameter μ of an irreducible highest weight module π_μ of G, in order for π_μ to be unitarizable. We have obtained this under the assumption that π_μ has a nonsingular infinitesimal character. We gather below the basic notation introduced in the course of our proof.

We denote by r the Borel subalgebra of g defined by the positive system P. For a parabolic subalgebra q of g containing r, we denote by m the unique reductive part of q containing the Cartan subalgebra b and call it the reductive part of q. Let P_m be the set of roots in P which are roots of the reductive part of q.

Theorem A. Let π_μ be an irreducible highest weight module for G which has highest weight μ (with respect to $P_k \cup -P_n$). Suppose that the infinitesimal character of π_μ is nonsingular. Now assume that π_μ is unitarizable. Then there exists a parabolic subalgebra q of g containing r such that

$$\mu = \lambda + 2\delta_{q,n}$$

where (i) $\delta_{q,n}$ is half the sum of the non-compact roots in the unipotent radical of q, (ii) $2(\lambda, \alpha)/(\alpha,\alpha)$ is a non-negative integer for all α in P and (iii) $(\lambda, \alpha) = 0$ for every root α in the reductive part of q.

Proof: Let P' be the positive system on which $\mu - \delta_n + \delta_k$ is positive. Then

(3.48) $\quad \mu = \lambda + \delta_n + \delta'_n$

where λ is an integral linear form, dominant with respect to P' and where δ'_n is half the sum of the non-compact roots in P' (cf. (3.7)). Let P'_n be the set of non-compact roots in P'. For each $Y \subseteq P_n \cap -P'_n$, let q_Y be the intersection of all parabolic subalgebras q of g containing r, such that Y is contained in the set of roots in the reductive part of q. For certain subsets Y, the set $P_n \cap P'_n$ is contained in the set of roots in the unipotent radical of q_Y. Choose a maximal one with this property and call this parabolic subalgebra q.

(3.49) For this q, we claim $P_n \cap P'_n$ is precisely the set of noncompact roots in the unipotent radical of q.

If this were not the case, we obtained a contradiction to the property (2.8) of unitarizable representations. Namely, we obtained (3.34). Thus, the assertion (3.49) is proved. It is therefore clear that $\delta_n + \delta'_n = 2\delta_{q,n}$. Hence by (3.48) $\mu = \lambda + 2\delta_{q,n}$. It remains to show the property, (ii) and (iii) for λ.

In (3.11), we proved that $(\lambda, \alpha) = 0$ for every α in $P_n \cap -P'_n$. Because of (3.49) $P_n \cap -P'_n$ is precisely the set of noncompact roots in P_m. We also know that the reductive part of q has no semisimple ideals contained in k (cf. (3.13)). Thus every compact root of m is a linear combination of noncompact roots in P_m. Thus $(\lambda, \alpha) = 0$ for every root α in P_m. This proves (iii) Note that

$$P = P_n \cup (P_n \cap P'_n) \cup (P_n \cap -P'_n).$$

Since λ is dominant with respect to P' and since P_k as well as $P_n \cap P'_n$ are contained in P', $(\lambda, \alpha) \geqslant 0$ if $\alpha \in P_k \cup (P_n \cap P'_n)$. If $\alpha \in P_n \cap -P'_n$, then we already saw (cf. (3.11)) that $(\lambda, \alpha = 0)$. Thus (ii) is proved. This completes the proof of theorem A.

In the next two sections, we will see that the converse of Theorem A is also true.

4. The sufficiency of the condition

The purpose of this section and the next one is to prove the following theorem, which is converse to theorem A.

Theorem B. Let q be a parabolic subalgebra of g containing r. Let $\delta_{q,n}$ be half the sum of the non-compact roots in the unipotent radical of q. Let λ be a linear form such that $2(\lambda, \alpha)/(\alpha, \alpha)$ is a nonnegative integer for every α in P and such that $(\lambda, \alpha) = 0$ for every root α in the reductive part of q. Let $\mu = \lambda + 2\delta_{q,n}$. Then the highest weight module (π_μ, H_μ) is unitarizable.

(4.1) *Remark.* π_μ, as in Theorem B, will have a nonsingular infinitesimal character.

The first part in the proof of Theorem B is the following.

(4.2) *Proposition.* Let (π_μ, H_μ) be as in Theorem B. Let L be the spin module for k. Let ξ be the highest weight of an irreducible k submodule of $H \otimes L$. Then $(\xi + \delta_k, \xi + \delta_k) \geqslant (\mu - \delta_n + \delta_k, \mu - \delta_n + \delta_k)$. Moreover, if V_ϕ is an irreducible k submodule of H_μ with highest weight $\phi \neq \mu$ and if V_ξ is an irreducible k submodule of $V_\phi \otimes L$ with highest weight ξ, then we actually have strict inequality $(\xi + \delta_k, \xi + \delta_k) > (\mu - \delta_n + \delta_k, \mu - \delta_n + \delta_k)$.

Proof: Our idea in proving (4.2) is to use the construction [5, § 4] where one builds a chain of $U(g)$ modules above g- Verma modules and takes a quotient of the biggest object of the chain to obtain modules like H_μ. We explain this a little more now.

For the discussion below, the condition that the positive system P is adapted to the complex structure on G/K is not needed. In fact G/K need not even admit any invariant complex structure and P could be arbitrary.

For the parabolic subalgebra q of g, let P_u be the set of roots in the unipotent radical of q and let P_m be the set of elements in P which are roots of the reductive part of q. Thus $P = P_m \cup P_u$ (disjoint). Also it is known that $(-P_m) \cup P_u$ is also a positive system for the roots of g. Let σ be the unique element of the Weyl group of g such that $\sigma P = (-P_m) \cup P_u$. For a while, let η be any regular integral linear form dominant with respect to $(-P_m) \cup P_u$. Set $W_1 = V_{g,P,-\eta-\delta}$. Let X equal the set of all simple roots of P which are elements of P_m. (Thus, X is simply the set of all simple roots of P_m). For each $\alpha \in X$, $2(-\eta - \delta, \alpha)/(\alpha, \alpha)$ is a non-negative integer. Hence for each $\alpha \in X$, the Verma module $V_{g,P,-s_\alpha\eta-\delta}$ is a proper submodule of $V_{g,P,-\eta-\delta}$. In fact, if we set W_0 equal to the sum $\Sigma_{\alpha\in X} V_{g,P,-s_\alpha\eta-\delta}$, then W_0 is a proper submodule of $W_1 = V_{g,P,-\eta-\delta}$. In the construction of [5] one builds a (finite) canonical chain of $U(g)$ modules containing W_1. The maximal object of this chain has a unique irreducible quotient. Let us here call it D_η. In [5] it is shown that D_η is a k-finite $U(g)$ module. From the work of [1] and [5] the module D_η, among other properties, has the following properties relating it to $V_{g,P,-\eta-\delta}$.

Let $P_{m,k}$ (resp $P_{u,k}$) be the compact roots in P_m (resp P_u). Then $P_k = P_{m,k} \cup P_{u,k}$. One knows that $(-P_{m,k}) \cup P_{u,k}$ is also a positive system for the roots of k.

Let τ (resp. t) be the unique element of the Weyl group W_k of k (regarded as a subgroup of the Weyl group of g) such that $\tau P_k = (-P_{m,k}) \cup P_{u,k}$ (resp. $tP_k = -P_k$), For $s \in W_k$, set $s'\phi = s(\phi + \delta_k) - \delta_k$. In [5] it is shown that

(4.3) $(t\tau)'(-\eta - \delta)$ is the highest weight of a k submodule of D_η, with multiplicity one.

(4.4) If ϕ is the highest weight of any k submodule of D_η, then ϕ is of the form $\phi = (t\tau)'(-\eta - \delta - A)$, where A is a nonnegative integral linear combination of elements of P and in addition $-\eta-\delta-A$ is a P_k extreme weight (cf. [1, § 2]) of W_1/W_0.

For (4.3), see [5, § 5] and for (4.4) see [5, Prop. 4.4].

(4.5) *Now suppose that* $(-\eta - \delta, a) = 0$ for every a in X.

Let u^- be the span of the root spaces not contained in q. Then

$$(4.6) \qquad g = u^- \oplus m \oplus u$$

where m (resp. u) is the reductive part of q (resp. the unipotent radical of q). If $C_{-\eta-\delta}$ is the one-dimensional weight space of $W_1/_{W_0}$ with weight $-\eta - \delta$, then the condition that $(-\eta - \delta, a) = 0$ for every a in X ensures that $u \cdot C_{-\eta-\delta} = 0$ and $m\, C_{-\eta-\delta} \subseteq C_{-\eta-\delta}$. Let $U(u^-)$, $U(m)$ and $U(u)$ denote respectively the enveloping algebras of u^-, m and u. Then $U(m) \cdot U(u) \cdot C_{-\eta-\delta} = C_{-\eta-\delta}$. Thus (4.6) implies $U(u^-) \cdot C_{-\eta-\delta} = W_1/_{W_0}$. In particular, any weight of $W_1/_{W_0}$ is of the form $-\eta-\delta-A$, where A is a non-negative integral linear combination of elements of P_u, the roots which occur in the unipotent radical of q.

Applying these remarks to (4.4) we obtain the following.

(4.7) Suppose $(-\eta - \delta, a) = 0$ for every a in X. If ϕ is the highest weight of any k submodule of D_η, then ϕ is of the form $(t\tau)'(-\eta - \delta - A)$ where A is a nonnegative integral linear combination of elements in P_u.

Now let V_ϕ be an irreducible k submodule of D_η with highest weight ϕ and let $\phi = (t\tau)'(-\eta - \delta - A)$ as in (4.7). Let V_ξ be an irreducible k submodule of $V_\phi \otimes L$ with highest weight ξ.

We wish to conclude that

$$(4.8) \qquad (\xi + \delta_k,\ \xi + \delta_k) \geqslant (\eta, \eta).$$

Let ψ be the lowest weight of an irreducible component of L. It is enough to prove the inequality (4.8) when ξ is in the W_k orbit of $\phi + \psi$. We know ψ is of the form $t\tilde{\delta}_n$, where $\tilde{\delta}_n$ is half the sum of the non-compact positive roots of some positive system $\tilde{P}$ for g such that $P_k \subseteq \tilde{P}$. Also, since ξ is dominant with respect to P_k and lies in the orbit of $\psi + t\tilde{\delta}_n$, it can be shown that for any $w \in W_k$,

$$(\xi + \delta_k,\ \xi + \delta_k) \geqslant (w(\phi + t\tilde{\delta}_n) + \delta_k,\ w(\phi + t\tilde{\delta}_n) + \delta_k).$$

Thus to prove (4.8) it suffices to show that for some $w \in W_k$,

$$(4.9) \qquad (w(\phi + t\tilde{\delta}_n) + \delta_k,\ w(\phi + t\tilde{\delta}_n) + \delta_k) \geqslant (\eta, \eta).$$

Now

$$\phi = (t\tau)'(-\eta - \delta - A) = t\tau(-\eta - \delta - A + \delta_k) - \delta_k$$
$$= t\tau(-\eta - \delta - A) + t(\tau\delta_k + \delta_k).$$

We will show (4.9) for the element $w = t\tau t$.

$$t\tau t\phi = t(-\eta - \delta - A) + t(\delta_k + \tau\delta_k).$$

Thus,

$$t\tau t(\phi + t\tilde{\delta}_n) = t(-\eta - \delta - A) + t(\delta_k + \tau\delta_k) + t\tau\tilde{\delta}_n.$$

So,

(4.10) $\quad t\tau t(\phi + t\tilde{\delta}_n) + \delta_k = t(-\eta - \delta - A) + t\tau\delta_k + t\tau\tilde{\delta}_n.$

In view of (4.10) to show (4.9) it is enough to show

(4.11) $\quad (-\eta - \delta - A + \tau\delta_k + \tau\tilde{\delta}_n,\ -\eta - \delta - A + \tau\delta_k + \tau\tilde{\delta}_n) \geqslant (\eta, \eta).$

Recall that $\tilde{\delta}_n$ was half the sum of the non-compact roots of a positive system $\tilde{P}$ such that $P_k \subseteq \tilde{P}$. Let $\triangle_m$ be the set of all roots of m. Then $\tilde{P} \cap \triangle_m$ gives a positive system $\tilde{P}_m$ for $\triangle_m$ and clearly $P_{m,k} \subseteq \tilde{P}_m$. Let $\tilde{P}^* =$ the positive system for the roots of g obtained by adjoining to $\tilde{P}_m$ the set P_u. Set $\tilde{\delta}_n^* =$ half the sum of the non-compact roots in $\tilde{P}^*$. Clearly,

(4.12) $\quad \tilde{\delta}_n = \tilde{\delta}_n^* - B$, where B is a sum of elements from $P_{u,n}$, the non-compact roots in P_u.

Also note that

(4.13) $\quad \tilde{P}_k \subseteq \tilde{P}^*$ and $P_u \subseteq \tilde{P}^*$.

Every element of the Weyl group of m leaves P_u stable. In particular $\tau P_u \subseteq P_u$. In view of these remarks, the two positive systems P and $\tau\tilde{P}^*$ both contain P_u and differ only in the roots of m. In particular if s is the unique element of the Weyl group of g such that $sP = \tau\tilde{P}^*$, then s is actually an element of the Weyl group of m. Hence

(4.14) $\quad sP_u = P_u$ and s can be written as a product of reflections s_α where the roots α are in X.

Note also that $s\delta = \tau\delta_k + \tau\tilde{\delta}_n^*$. We write it as

(4.15) $\quad s\delta = \delta + (-\delta + \tau\delta_k + \tau\tilde{\delta}_n^*)$

and think of it as the result obtained by applying the formula $s_\alpha(\lambda) = \lambda - 2(\lambda, \alpha)/(\alpha,\alpha)\ \alpha$ successively to the reflections s_α in the expression for s as in (4.14). Since $(-\eta, \alpha) = (\delta, \alpha)$ for every α in X, it follows from (4.15) that

(4.16) $\quad s(-\eta) = -\eta - \delta + \tau\delta_k + \tau\tilde{\delta}_n^*.$

With these preparations we can now show (4.11). Using (4.12) and (4.16), we see that

(4.17) $\quad -\eta-\delta-A+\tau\delta_k+\tau\tilde{\delta}_n=s(-\eta)-A-\tau B$

where A and τB are both nonnegative integral linear combinations of elements from P_u. We assumed that η was regular and dominant with respect to $(-P_m)\cup P_u$. Since $sP_u=P_u$, we see that $s(-\eta)$ is dominant with respect to $(sP_m)\cup(-P_u)$. Thus, from (4.17) we see that

$$(-\eta-\delta-A+\tau\delta_k+\tau\tilde{\delta}_n,\ -\eta-\delta-A+\tau\delta_k+\tau\delta_n)\geqslant(\eta,\eta)$$

and equality occurs only if $A=0$ and $B=0$.

Thus, we have shown (4.8) ; in fact we also proved that for equality to hold in (4.8), it is necessary that $\phi=(t\tau)'(-\eta-\delta)$, i.e., V_ϕ should be the unique 'minimal' k-type of D_η.

We now apply these general facts to our special case to prove proposition (4.2). First of all observe that for μ as in theorem B, the statements in proposition 4.2 for (π_μ, H_μ) will follow if we prove the corresponding statements for (π_μ^*, H_μ^*) the dual of (π_μ, H_μ). We will prove the statements for (π_μ^*, H_μ^*) by identifying H_μ^* with a 'D_η', described above. In fact, choose $\eta=\lambda-\delta_m+\delta_u$.
Then, we claim

(4.18) $\quad D_\eta\approx H_\mu^*$.

We will first verify

(4.19) $\quad (t\tau)'(-\eta-\delta)=-t\mu$

$$\begin{aligned}(t\tau)'(-\eta-\delta)&=t\tau(-\eta-\delta+\delta_k)-\delta_k\\&=t(\tau(-\eta-\delta+\delta_k)+\delta_k)\\&=t(\tau(-\eta-\delta_m-\delta_u+\delta_k)+\delta_k)\\&=t(\tau(-\lambda-2\delta_u)+\tau\delta_k+\delta_k)\\&=t(-\lambda-2\delta_u+\tau\delta_k+\delta_k)\end{aligned}$$

since $\tau P_u=P_u$ and τ is a product of reflections s_a, $a\in X$, and $(\lambda,a)=0$ for $a\in X$.
But $-2\delta_u+\tau\delta_k+\delta_k=-2\delta_{a,n}$.
So, $t(-\lambda-2\delta_u+\tau\delta_k+\delta_k)=t(-\lambda-2\delta_{a,n})=-t\mu$ and this proves (4.19).
Now, let ϕ be the highest weight of an irreducible k submodule of D_η. By (4.4) ϕ is of the form $\phi=(t\tau)'(-\eta-\delta-A)$ where A is nonnegative integral linear combination of elements of P. Here, in addition $-\eta-\delta-A$ should be a P_k extreme weight for W_1. If V_1 denotes the k—Verma module $V_{k,P_k,-\eta-\delta}$ contained in $W_1=V_{g,P,-\eta-\delta}$, then the action of the enveloping algebra gives a k module surjection $U(p)\otimes V_1\to W_1$. Hence any P_k extreme vector of W_1 has to be of the form $-\eta-\delta-A$ where A is a nonnegative integral linear combination of elements of P_n. Since elements of W_k leave P_n stable, we now conclude that ϕ is of the form $(t\tau)'(-\eta-\delta)-B$ where B is a nonnegative integral linear combination of elements of P_n. This shows that D_η^* is a highest weight module

(with respect to $P_k \cup -P_n$ as in Def. (1.1)) with highest weight $-t((t\tau)'(-\eta-\delta))$. Since $(t\tau)'(-\eta-\delta) = -t\mu$ (by (4.19) the assertion (4.18) follows.

To conclude the inequalities and the statements in proposition (4.2), it remains to verify $(-\eta-\delta, \alpha) = 0$ for α in X. But

$$-\eta-\delta = -\lambda+\delta_m-\delta_u-(\delta_m+\delta_u) = -\lambda-2\delta_u.$$

But by assumption $(\lambda, \alpha) = 0$, for α in X and also it is well known that $(2\delta_u, \alpha) = 0$ for α in X.

This completes the proof of proposition (4.2). In the next section, we will use the result of proposition (4.2) and arrive at the unitarizability of (π_μ, H_μ).

5. Role of the formal Dirac operator

The purpose of this section is to prove a general unitarizability result for highest weight modules when one knows an inequality as in proposition (4.2).

Let (π_μ, H_μ) be an irreducible highest weight module which admits an invariant hermitian form.

(5.1) *Proposition.* Let ϕ be the highest weight of an irreducible k submodule V_ϕ of H_μ and suppose that for every k submodule V_ξ contained in $V_\phi \otimes L$, one has $(\xi+\delta_k, \xi+\delta_k) \geqslant (\mu-\delta_n+\delta_k, \mu-\delta_n+\delta_k)$, with strict inequality whenever $\phi \neq \mu$. Then (π_μ, H_μ) is unitarizable.

The techniques of proving (5.1) are essentially the same as already employed in the proof of [4, Prop. 9.7)]. For the benefit of the reader, we will discuss below the main ingredients of that argument.

We define a filtration H_i in $H = H_\mu$ as follows. H_0 is the irreducible k module V_μ. We inductively define $H_{i+1} = H_i + p_+ H_i$. Since H is a highest weight module, $H = U_i H_i$. Now, let us normalize the hermitian forms on H, so that it restricts to a positive definite one on H_0. Inductively, assume that it restricts to a positive definite one on H_i. We wish to prove it restricts to a positive definite one on H_{i+1}.

The main tool we employ is the formal Dirac operator $D : H \otimes L \to H \otimes L$, (cf. (2.1)). Clearly $D(H_i \otimes L) \subseteq H_{i+1} \otimes L$. Let $L_{-\delta_n}$ be the one dimensional k submodule of L whose highest is $-\delta_n$.

(5.2) *Lemma.* $D(H_{i+1} \otimes L_{-\delta_n}) \subseteq H_i \otimes L$.

With suitable normalizations,

$$(5.3)\quad D = \Sigma_{\alpha \in P_n} \pi(X_\alpha) \otimes C(X_{-\alpha}) + \Sigma_{\alpha \in P_n} \pi(X_{-\alpha}) \otimes C(X_\alpha).$$

Clearly the part $\Sigma_{\alpha \in P_n} \pi(X_\alpha) \otimes C(X_{-\alpha})$ annihilates $H_{i+1} \otimes L_{-\delta_n}$. Hence to prove (5.2) it is enough to show that $p_- \cdot H_{i+1} \subseteq H_i$. This can be done easily. Thus (5.2) is proved.

Take the standard hermitian form on L. Then we have a product hermitian form on $H \otimes L$. One can show that if $v, w \in I_\xi \subseteq H \otimes L$, where I_ξ is the isotypical k submodule of $H \otimes L$ with highest weight ξ, then,

$$(5.4)\quad (Dv, Dw) = \{(\xi+\delta_k, \xi+\delta_k) - (\mu-\delta_n+\delta_k, \mu-\delta_n+\delta_k)\}(v, w).$$

Let H^+ be the unique k-submodule of H_i which is a complement of H_0. Because of the hypothesis of the proposition (5.1) the scalar within brackets in (5.4) is a positive number whenever $v, w \in (U_i H_i^+) \otimes L$. It now follows from (5.4) and (5.2) and the induction hypothesis that the hermitian form on $H_{i+1}^+ \otimes L_{-\delta_n}$ is positive definite. Dividing out by the + ve factor coming from $L_{-\delta_n}$, we see that on H_{i+1}^+ the form is + ve definite. Since H_0 and H_{i+1}^+ are orthogonal, it now follows that on H_{i+1} the form is +ve definite.

This proves proposition (5.1).

Now, combining together proposition (4.2) and proposition (5.1), we have proved theorem *B*, since the modules in question are known to possess invariant hermitian forms.

6. Applications to (o, p) Betti numbers : Remarks

Let Γ be a discrete subgroup of G so that $\Gamma \backslash G / K$ is a compact locally symmetric hermitian domain. The (o, p) Betti number of $\Gamma \backslash G / K$ is a certain sum over the class of irreducible unitary highest weight modules (π, H) for G, having the same infinitesimal character as the trivial one-dimensional representation of G (cf. [2]). If such a module (π, H) has a nonzero contribution to the (o, p) Betti number, then necessarily,

(6.1) $\dim (\mathrm{Hom}_k (\wedge^p p_+, H)) \neq 0.$

In particular, (π, H) has to be a module for the adjoint group of g_0. Moreover, since (π, H) has the same infinitesimal character as the trivial one-dimensional representation of G, the infinitesimal character of π is regular. Using theorem *A* and theorem *B*, we obtain the following.

(6.2) *Proposition.* Let q be a parabolic subalgebra of g containing r. Let $\mu = 2\delta_{q,n}$, the sum of all the non-compact roots in the unipotent radical of q. The highest weight modules (π_μ, H_μ) obtained this way for the various q consist precisely of the set of irreducible unitary highest weight modules for G having the same infinitesimal character as the trivial one-dimensional module.

(6.3) Let q be as in proposition (6.2) and let $\mu = 2\delta_{q,n}$. When is $\mathrm{Hom}_k (\wedge^p p_+, H_\mu)$ nonzero ? It is nonzero if and only if p is exactly the number of non-compact roots in the unipotent radical of q.

Suppose p is the number of non-compact roots in q and let $X_{\alpha_1}, \ldots, X_{\alpha_p}$ be the corresponding root vectors. The vector $X_{\alpha_1} \wedge \ldots \wedge X_{\alpha_p}$ in $\wedge^p p_+$ has weight μ. If β is a positive compact root, then for $1 \leqslant i \leqslant p$, either $[X_\beta, X_{\alpha_i}] = 0$ or else, $[X_\beta, X_{\alpha_i}]$ is a scalar multiple of X_{α_j}, $1 \leqslant j \leqslant p, j \neq i$. Hence $\mathrm{ad}(X_\beta)(X_{\alpha_1} \wedge \ldots \wedge X_{\alpha_p}) = 0$. Thus $X_{\alpha_1} \wedge \ldots \wedge X_{\alpha_p}$ is a highest weight vector with highest weight μ. This proves $\mathrm{Hom}_k (\wedge^p p_+, H_\mu) \neq 0$.

Conversely, suppose $\mathrm{Hom}_k (\wedge^p p_+, H_\mu) \neq 0$. Then there exists an irreducible k module V_ϕ with highest weight ϕ such that $V_\psi \subseteq \wedge^p p_+$ and $V_\phi \subseteq H_\mu$. Since $\wedge^p p_+ \subseteq L \otimes L^*$, we have $\mathrm{Hom}_k (V_\phi \otimes L, L) \neq 0$. Hence we can find an irreducible k module V_ξ with highest weight ξ such that $V_\xi \subseteq V_\phi \otimes L$ and $V_\xi \subseteq L$. Since $V_\xi \subseteq L$, $(\xi + \delta_k, \xi + \delta_k) = (\delta, \delta)$ (cf. [3, § 2]).

Note that since H_μ has the same infinitesimal character as the trivial one-dimensional module, $(\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k) = (\delta, \delta)$.

Thus $(\xi + \delta_k,\ \xi + \delta_k) = (\mu - \delta_n + \delta_k,\ \mu - \delta_n + \delta_k)$. As we already have $V_\xi \subseteq V_\phi \otimes L$ and $V_\phi \subseteq H_\mu$, we conclude from proposition (4.2) that $V_\phi = V_\mu$. Thus, $\mathrm{Hom}_k(\wedge^p p_+,\ V_\mu) \neq 0$. But if s is the number of non-compact roots of q, then as we already saw $\mathrm{Hom}_k(\wedge^s p_+,\ V_\mu) \neq 0$. Thus, $\mathrm{Hom}_k(\wedge^p p_+,\ \wedge^s p_+) \neq 0$. This can happen only if $p = s$. Thus (6.3) is proved.

Since the multiplicity of V_ϕ in $\wedge^p p_+$ can be at most one, in the course of the above argument, we have actually proved

(6.4) The space $\mathrm{Hom}_k(\wedge^p p_+,\ H_\mu)$ in (6.3) has dimension exactly one.

The Betti numbers of $\Gamma \backslash G / K$, have been studied through representation theory by Matsushima, Hotta–Wallach, Borel–Wallach, Zuckerman and Casselman-Schmid. In particular if r is the real rank of G and if $1 \leqslant p < r$, their results show that the (o, p) Betti number must be zero. Combining this with our observations in this section, we should expect that if $1 \leqslant p < r$, there does not exist any parabolic subalgebra containing the Borel subalgebra r for which p is the number of non-compact roots in the unipotent radical of q. In fact, when we set out to verify this, case by case, we see that this is always the case; occasionally (e.g. $SO^\circ(n, 2)$, $SO^*(2n)$ and the exceptionals) we even get sharper results. We list below the result of doing this exercise.

(6.5) $G = SU(m, n)$ $(m \geqslant n)$. Real rank $= n$.

In the Dynkin diagram there are $m + n - 1$ vertices $a_1, a_2, \ldots, a_{m+n-1}$ enumerated in the 'usual' way. The unique non-compact root is a_m. Let q be the maximal parabolic subalgebra defined by $(a_2, a_3, \ldots, a_{m+n-1})$. The cardinality of $P_{u,n}$ (non-compact roots in the unipotent radical of q) is n. There is no parabolic subalgebra q containing r for which $1 \leqslant \# P_{u,n} < n$.

(6.6) $G = SO^*(2n)$ $(n > 3)$. Real rank $= r = [\frac{1}{2}n]$.

The Dynkin diagram has vertices $a_1, \ldots, a_n$ with a_n and a_{n-1} forming a wedge at a_{n-2}. The unique non-compact root is a_n. Let q be the maximal parabolic subalgebra defined by $(a_2, \ldots, a_{n-1}, a_n)$. Then the cardinality of $P_{u,n}$ is $n - 1$. There is no other q containing r for which, $1 \leqslant \# P_{u,n} \leqslant n - 1$. The (o, p) Betti numbers in this case vanish for $1 \leqslant p < n - 1$, even though real rank is $[\frac{1}{2}n]$.

(6.7) $G = SO(n, 2)$ $(n > 2)$. Real rank $= 2$. Let $(a_1, a_2, \ldots)$ be the vertices in the Dynkin diagram. Any possible wedge (which only occurs if n is even) is supposed to be at the right end. The unique non-compact simple root is a_1. Let q be the maximal parabolic subalgebra defined by omitting the last simple root. Then cardinality of $P_{u,n} = [(n + 1)/2]$, the integral part of $[(n + 1)/2]$. There is no parabolic subalgebra q containing r for which $1 \leqslant \# P_{u,n} < [(n + 1)/2]$. Hence, the (o, p) Betti numbers vanish for $1 \leqslant p < [(n + 1)/2]$.

(6.8) $G = Sp(n, R)$. Real rank $= n$. The vertices in the Dynkin diagram are $(a_1, a_2, \ldots, a_n)$ and a_n is the unique non-compact simple root. Let q be the

maximal parabolic subalgebra of g defined by $(\alpha_2, \ldots, \alpha_n)$. Then cardinality of $P_{u,n}$ is n. There is no other parabolic subalgebra q containing r for which $1 \leqslant \# p_{u,n} \leqslant n$. Therefore, in this case, we do not get vanishing of (o, p) Betti numbers sharper than those already known.

(6.9) $G =$ the unique real form of E_6, whose symmetric space is hermitian. The real rank is 2. The dimension of p_+ is 16. The Dynkin diagram has vertices $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6)$ where the part $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ is of type A_5 and α_6 is connected to α_3. The unique non-compact simple root is α_1. Let q be the parabolic subalgebra of g defined by omitting α_5. The cardinality of $P_{u,n}$ is 8. There is no other parabolic subalgebra q containing r of g for which $1 \leqslant \# P_{u,n} \leqslant 8$. Thus, in this case the (o, p) Betti numbers vanish for $1 \leqslant p < 8$ (even though real rank is 2). As q varies the set of numbers $\# P_{u,n}$ that we get is precisely (0, 8, 11, 12, 13, 14, 15, 16). Thus, the (o, p) Betti numbers vanish also for $p = 9$ and $p = 10$.

(6.10) $G =$ the unique real form of E_7, whose symmetric space is hermitian. The real rank is 3. The dimension of p_+ is 27. The set of numbers $P_{u,n}$ as q (containing r) varies is precisely (0, 17, 21, 22, 23, 24, 25, 26, 27). Thus the (o, p) Betti numbers vanish for $1 \leqslant p < 17$ and for $p = 18, 19, 20$.

These cases cover all irreducible hermitian symmetric spaces.

(6.11) *Remark.* In the case of $G = Sp(n, R)$, the numbers $\# P_{u,n}$ as q varies consist precisely of the set $\{0\} \cup \{n + (n-1) + \ldots + (n-i) \mid i = 0, 1, 2, \ldots, n-1\}$. Thus if p does not belong to this set the (o, p) Betti number is zero. In the case of $G = SU(m, n)$, the set of numbers $\# P_{u,n}$ is precisely $\{mn - m'n' \mid o \leqslant m' \leqslant m,\ o \leqslant n' \leqslant n,\ m' \text{ and } n' \text{ are integers}\}$ and so the (o, p) Betti numbers vanish if p is not in this set. Similar descriptions can be obtained for the other cases also.

References

[1] Enright T J and Varadarajan V S 1975 *Ann. Math.* **102** 1
[2] Hotta R and Wallach N R 1975 *Osaka J. Math.* **12** 419
[3] Parthasarathy R 1972 *Ann. Math.* **96** 1
[4] Parthasarathy R 1977 *Math. Ann.* **226** 1
[5] Parthasarathy R 1978 *Comp. Math.* **36** 53
[6] Parthasarathy K R, Ranga Rao R and Varadarajan V S 1967 *Ann. Math.* **85** 383
[7] Schmid W 1975 *Inventiones Math.* **30** 47

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 25–33.

Nonnegative integral solution of linear equations

S K SEN
Computer Centre, Indian Institute of Science, Bangalore 560 012

MS received 7 September 1978

Abstract. A method to obtain a nonnegative integral solution of a system of linear equations, if such a solution exists is given. The method writes linear equations as an integer programming problem and then solves the problem using a combination of artificial basis technique and a method of integer forms.

Keywords. All-integer programming; artificial basis technique; Gomory method of integer forms; linear programming; non-negative integral solution.

1. Introduction

Many problems like those of path-length, fixed-charge, batch-size, transportation and allocation, chemical reactor vessel (Gottfried and Weisman [7]), computer networking involve quantities which can be only nonnegative integers. Such a problem giving rise to linear equations needs nonnegative integral solution.

Hurt and Waid [8] propose a generalized inverse A^- which gives the general integral solution (all the integral solutions) to linear equations (also Ben-Israel and Greville [1]; Marcus and Minc [9]; Sen and Shamim [12]). There seems to be no easy way to seeve out nonnegative integral solutions from the general form $x = A^-b + (I - A^-A)y$ where y and I are arbitrary integral n-vector and $n \times n$ unit matrix, respectively. $Ax = b$ are the equations where A is an $m \times n$ integral matrix.

The method described here investigates equations $Ax = b$, consistent or not, underdetermined or overdetermined as an all-integer programming (all-ip) problem and gives a nonnegative integral solution x when it exists. To solve the all-ip problem the method involves a particular form of the artificial basis technique (Sen [11]; Chung [2]; Strum [13]) and the Gomory method of integer forms (Gomory [6]; Vajda [15]; Salkin [10]).

2. The problem

Obtain a nonnegative integral solution x of $Ax = b$ (if it exists) where $A = (a_{ij})$ is a given $m \times n$ integral matrix, $b = (b_i)$ is a given non-negative integral m-vector and $x = (x_j)$ is an n-vector. (1)

Note. There is no loss of generality in considering (*i*) $b \geqslant 0$ and (ii) A and b integral. For, if these are not then multiply the equations with negative b_i by -1 and nonintegral equations by suitable scalars.

3. Existence of a nonnegative integral solution

$Ax = b$ has a non-negative solution x if and only if $A^t y \geqslant 0$, $b^t y < 0$ has no solution y. t indicates the transpose (Vajda [14]; Farkas [3]).

Equivalently, $Ax = b$ has no nonnegative solution x if and only if $A^t y \geqslant 0$, $b^t y < 0$ has a solution y.

Let A and b be integral and $Ax = b$ be consistent (i.e., $AA^- b = b$). Also, let $Ax = b$ have a nonnegative solution. If A is nonsingular and its inverse is also integral then $Ax = b$ has unique nonnegative integral solution $x = A^{-1} b$. Further, if $x = A^- b$ is not integral then $Ax = b$ has no integral solution (it has non-integral rational solutions though). A^- here is a (reflexive) generalized inverse that satisfies

$$AA^- A = A,\ A^- AA^- = A^-,\ A^- A \text{ and } AA^- \text{ are integral.}$$

These results are not of immediate use. However, the method tells if a nonnegative solution of $Ax = b$ does not exist. In fact, the necessary and sufficient condition for $Ax = b$ to have a nonnegative solution is the method producing one. Further, the sufficient condition that this solution is integral is the (Gomory) method giving one.

4. The method

The method consists of two parts.

Part 1 (*Equivalent ip problem*). Write (1) as an all-ip problem.

Part 2 (*Gomory-artificial-basis technique*). Solve this ip problem using Gomory method of integer forms in which a particular form of the artificial basis technique is embedded.

(i) *Equivalent ip problem*

Let x and A be now extended $(n + m)$-vector and $m \times (n + m)$ matrix, respectively. Further, let the last m columns of A form an $m \times m$ unit matrix. The *ip* problem equivalent to (1) is

Obtain x so that

Min $z = x_{n+1} + \ldots + x_{n+m} = 0$: Objective function

subject to

$Ax = b$: Constraints

$x \geqslant 0$ and integral : Nonnegativity and integrality conditions. (2)

(ii) *Gomory-artificial-basis technique*

Step 1. Solve the *ip* problem as a linear programming (*lp*) problem using the artificial basis technique in 'restricted tableau' (described later). If it is infeasible,

so is the *ip* problem—terminate. If the optimal solution is all integer then the *ip* problem is solved—terminate. Otherwise go to step 2.

Step 2. Consider one of the variables which have a fraction* in their value in the optimal (simplex) restricted tableau.

(i) Let the row of such a variable be

$$x \quad z_1 \quad .. \quad z_t \quad z_0$$

where z_0 is the present value of the variable x.

(ii) Write every z_i as $L_i + f_i$, where $L_i =$ the largest integer contained in z_i and hence f_i is nonnegative. In particular, f_0 is positive.

(iii) Add to the (simplex) tableau the further row (Gomory constraints)

$$s_1 \quad -f_1 \quad .. \quad -f_t \quad -f_0.$$

(iv) Apply the Dual Simplex method (described later) on this tableau. This renders the new variable s_1 non-basic.

Note. The Simplex tableau is already dual feasible, since the final tableau of the artificial basis technique (Simplex method) is reached. So the Dual Simplex method has been used.

Step 3. If the result again contains a basic variable which is not an integer then continue introducing new variables, $s_2, \ldots$ The method terminates in a finite number of steps if the feasible region of the *ip* problem is bounded (sufficient but not necessary).

5. Artificial basis technique in restricted tableau

Step 1. Set up the restricted Simplex tableau for (2), and write the coefficients (in parentheses) which x_j have in the objective function and the last row, i.e., d_j-row using the checking rule (described later) as below

$$\begin{array}{lccccc}
 & (0) & .. & (0) & .. & (0) & \\
 & x_1 & & x_{j_0} & & x_n & \mathrm{b} \\
(1)\ x_{n+1} & a_{11} & & a_{1j_0} & & a_{1n} & b_1 \\
 & \vdots & & & & & \\
(1)\ x_{n+i_0} & a_{i_0 1} & & a_{i_0 j_0} & & a_{i_0 n} & b_{i_0} \\
 & \vdots & & & & & \\
(1)\ x_{n+m} & a_{m1} & & a_{mj_0} & & a_{mn} & b_m \\
 & d_1 & & d_{j_0} & & d_n & d_{n+1}
\end{array} \qquad (3)$$

Step 2. (pivot selection). Let d_{j_0} be positive. Consider then, for all positive a_{ij_0}, the ratios b_i/a_{ij_0} and take a smallest. If this is obtained for i_0 then call $p = a_{i_0 j_0}$

* It is generally assumed that convergence is speeded by choosing that cut which bites as deeply as possible. This is usually taken to mean the selection of the row that gives the largest fraction (f_0).

the pivot (marked with a plus). Go to Step 3. Otherwise the present tableau is final and it either indicates no solution of $Ax = b$ or gives a nonnegative solution.

Step 3 (next-tableau computation). Having interchanged x_{j_0} and x_{n+i_0} obtain the next tableau as follows.

$$\begin{array}{lcccc} & x_1 & \cdots & x_{n+i_0} & \cdots & x_n & b \\ x_{n+1} & & & -a_{ij_0}/p & & & \\ & & & \vdots & & & \\ x_{j_0} & a_{i_0 1}/p & & 1/p & & a_{i_0 n}/p & b_{i_0}/p \\ & & & \vdots & & & \\ x_{n+m} & & & -a_{mj_0}/p & & & \\ & & & -d_{j_0}/p & & & \end{array} \tag{4}$$

The blank positions are filled in as follows :

$$\begin{aligned} a_{ij} &\leftarrow a_{ij} - a_{i_0 j} a_{ij_0}/p \\ d_j &\leftarrow d_j - a_{i_0 j} d_{j_0}/p. \end{aligned} \tag{5}$$

Note. '$\leftarrow$' means 'is replaced by'.

(i) The foregoing two 'replacements' are actually identical when we consider the last row (i.e., d_j-row) as just another row like the rows of (a_{ij}).

(ii) The right-hand side elements are the elements of the foregoing tableau throughout the computation.

Step 4 (termination condition). If the bottom row i.e. d_j-row excluding the last element is nonpositive, or if none of $x_{n+1}, \ldots, x_{n+m}$ occurs in the basis with a nonzero value then a nonnegative solution is reached. Otherwise go to step 2.

6. The checking rule for a simplex tableau

Let the *lp* problem be

Minimize $z = c^t x$ subject to $Ax = y$, $x \geqslant 0$

where

$$c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \; b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}, \; A = \begin{bmatrix} a_{11} \cdots a_{1n} \\ \vdots \quad \vdots \\ a_{m1} \cdots a_{mn} \end{bmatrix}.$$

We attach to all variables x_i the coefficient which they have in the objective function. Let, for example, a current 'restricted tableau' be

$$\begin{array}{lccccc} & (c_1) & (c_6) & (c_2) & (c_4) & \\ & x_1 & x_6 & x_2 & x_4 & \\ (c_3)\; x_3 & p_{11} & p_{12} & p_{13} & p_{14} & v_1 \\ (c_5)\; x_5 & p_{21} & p_{22} & p_{23} & p_{24} & v_2 \\ & d_1 & d_2 & d_3 & d_4 & d_5 \end{array}$$

Then

$$\begin{aligned} c_3p_{11} + c_5p_{21} - c_1 &= d_1 \\ c_3p_{12} + c_5p_{22} - c_6 &= d_2 \\ c_3p_{14} + c_5p_{24} - c_2 &= d_3 \\ c_3v_1 + c_5v_2 &= d_5. \end{aligned}$$

Such a relationship holds in all tableaux. This relationship is referred to as the *checking rule* for a tableau. Satisfaction of this rule is necessary for a restricted tableau to be correct but it is not sufficient (i.e., the rule may be satisfied even if a computational mistake occurs).

7. The dual simplex method

When to use. Let an *lp* problem be

$$\begin{aligned} &\text{Minimize} && z = c_1x_1 + \cdots + c_nx_n \\ &\text{subject to} && a_{11}x_1 + \cdots + a_{1n}x_n + a_{1\,n+1}\,x_{n+1} = b_1 \\ &&& \vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ &&& a_{m1}x_1 + \cdots + a_{mn}x_n + a_{mn+1}\,x_{n+m} = b_m \\ &&& x_i \geqslant 0 \quad i = 1\,(1)\,n+m. \end{aligned}$$

Also, let

$$a_{1n+1} = \cdots = a_{mn+1} = -1$$

and all c_j $j = 1\,(1)\,n$ be non-negative so that, in the first tableau, the first n elements in the bottom row are nonpositive (since we minimize). We call such a tableau *dual feasible.* If, in addition, all b_i $i = 1\,(1)\,m$ are nonnegative then the result is reached. Otherwise apply dual simplex method.

The method

Step 1 (pivot selection). Let b_{i_0} be negative. Consider, for all $a_{i_0j} < 0$, $|\,c_j/a_{i_0j}\,|$ and take a smallest. If this is obtained for j_0 then $a_{i_0j_0}$ is the pivot.

Step 2. (next-tableau computation). Same as in the Simplex algorithm (Vajda [15], Chung [2]; Gass [4]) or as in Step 3 of Sec. 5.

Step 3 (termination condition). If the bottom row (i.e., c_j-row) excluding the last element is nonpositive then the solution is reached—terminate. Otherwise go to step 1.

8. Examples

(i) Obtain a nonnegative integral solution of

$$\begin{bmatrix} -1 & 5 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

Restricted tableau 0

	(o)	(o)	(o)	(o)	
	x_1	x_2	x_3	x_4	b
(1) x_5	−1	5+	1	0	5
(1) x_6	1	2	0	1	8
	0	7	1	1	13

Restricted tableau 1

	x_1	x_5	x_3	x_4	b
x_2	−1/5	1/5	1/5	0	1
x_6	7/5+	−2/5	−2/5	1	6
	7/5	−7/5	−2/5	1	6

Restricted tableau 2

	x_6	x_5	x_3	x_4	b
x_2	1/7	1/7	1/7	1/7	13/7
x_1	5/7	−2/7	−2/7	5/7	30/7
	−1	−1	0	0	0

A nonnegative solution is thus $x = (30/7 \quad 13/7 \quad 0 \quad 0)^t$.

Consider the first row as it contains the largest fraction in the value of the variable x_2, viz., 6/7. Generate the new row (Gomory constraint) as in the algorithm and append this row. Thus

Restricted tableau 20

	x_6	x_5	x_3	x_4	b
x_2	1/7	1/7	1/7	1/7	13/7
x_1	5/7	−2/7	−2/7	5/7	30/7
s_1	−1/7	−1/7	−1/7+	−1/7	−6/7
	−1	−1	0	0	0

Restricted tableau 21

	x_6	x_5	s_1	x_4	b
x_2	0	0	1	0	1
x_1	1	0	−2	1	6
x_3	1	1	−7	1	6
	−1	−1	0	0	0

Hence a nonnegative integral solution is $x = (6 \quad 1 \quad 6 \quad 0)^t$.

Note. $x = (0 \quad 1 \quad 0 \quad 6)^t$ could also be another nonnegative integral solution.

(ii) *Degenerate case* (redundant equation)

Obtain a nonnegative integral solution (Sen [17]) of

$$\begin{bmatrix} -1 & 2 & 3 & 3 \\ 2 & 5 & 6 & 3 \\ -5 & -8 & -9 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 7 \\ 16 \\ -25 \end{bmatrix}$$

Restricted tableau 0

	(o) x_1	(o) x_2	(o) x_3	(o) x_4	b
(1) x_5	−1	2	3+	3	7
(1) x_6	2	5	6	3	16
(1) x_7	5	8	9	3	25
	6	15	18	9	48

Restricted tableau 1

	x_1	x_2	x_5	x_4	b
x_3	−1/3	2/3	1/3	1	7/3
x_6	4	1+	−2	−3	2
x_7	8	2	−3	−6	4
	0	3	−6	−9	6

Note. The last equation has been multiplied by −1 to make b_3 positive.

Restricted tableau 2

	x_1	x_6	x_5	x_4	b
x_3	− 3	−2/3	5/3	3	1
x_2	4	1	−2	−3	2
x_7	0	−2	1	0	0
	−12	−3	0	0	0

The artificial variable x_7 remains in the basis with a zero value. A nonnegative integral solution is $x = (0 \quad 2 \quad 1 \quad 0)^t$.

(iii) *Nonnegative nonintegral solution*

Obtain a nonnegative integral solution of

$$\begin{bmatrix} 1 & 2 & 1 \\ -4 & -2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Restricted tableau 0

	(o) x_1	(o) x_2	(o) x_3	b
(1) x_4	1	2	1	1
(1) x_5	−4	−2	3+	2
	−3	0	4	3

Restricted tableau 1

	x_1	x_2	x_5	b
x_4	7/3	8/3+	−1/3	1/3
x_3	−4/3	−2/3	1/3	2/3
	7/3	8/3	−4/3	1/3

Restricted tableau 2

	x_1	x_4	x_5	b
x_2	7/8	3/8	−1/8	1/8
x_3	−3/4	2/8	1/4	3/4
	0	−1	−1	0

Hence a nonnegative solution is $\mathbf{x} = (0 \quad 1/8 \quad 3/4)^t$.

Adding the Gomory constraint we have

Restricted tableau 20

	x_1	x_4	x_5	b
x_2	7/8	3/8	−1/8	1/8
x_3	−3/4	1/4	1/4	3/4
s_1	−1/4+	−1/4	−1/4	−3/4
	0	−1	−1	0

Restricted tableau 21

	s_1	x_4	x_5	b
x_2	7/2	−1/2	−1+	−5/2
x_3	−3	1	1	3
x_1	−4	1	1	3
	0	−1	−1	0

	s_1	x_4	x_2	b
x_5	−7/2	1/2	−1	5/2
x_3	1/2	1/2	1	1/2
x_1	−1/2	1/2	1	1/2
	−7/2	−1/2	−1	5/2

By adding the Gomory constraint we obtain the last row except the last element (viz., 5/2) nonpositive and one artificial variable, viz., x_5 is still in the basis with the nonzero value 5/2. Hence the equations have no integral solution.

(iv) *Inconsistent equations*

Obtain a nonnegative integral solution (Sen [11]) of

$$\begin{bmatrix} 5 & 3 & 2 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix}$$

Restricted tableau 0

	(o)	(o)	(o)	
	x_1	x_2	x_3	b
(1) x_4	5	3	2	10
(1) x_5	2	1	2	5
(1) x_6	4	2+	4	1
	11	6	8	16

	x_1	x_6	x_3	b
x_4	−1	−3/2	−4	17/2
x_5	0	1	0	9/2
x_2	2	1/2	2	1/2
	−1	−3	−4	13

The last row except the last element (viz, 13) is nonpositive and two artificial variables, viz. x_4 and x_5 are still in the basis with nonzero values. Hence the equations have no nonnegative solution. In fact, the equations have no solution at all.

Acknowledgement

The author wishes to thank Dr A A Shamim, Chairman, Computer Centre, Indian Institute of Science for constant encouragement.

References

[1] Ben-Israel A and Greville T N E 1974 *Generalized inverses : Theory and applications* (New York : Wiley-Interscience)
[2] Chung A 1966 *Linear Programming* (Columbus, Ohio: Charles, E. Merrill Books, Inc.)
[3] Farkas J 1901 *J.r. Angew. Math.* **124** 1
[4] Gass S I 1975 *Linear programming : Methods and applications* (New York : McGraw-Hill)
[5] Gomory R 1963 '*All-integer integer programming algorithm*' *in Industrial Scheduling* (eds Muth and Thompson) (Englewood Cliffs., New Jersey : Prentice-Hall)
[6] Gomory R 1963 *Recent advances in mathematical programming* (eds Graves and Wolfe), (New York : McGraw-Hill)
[7] Gottfried B S and Weisman J 1973 *Introduction to optimization theory* (Englewood Cliffs, New Jersey : Prentice-Hall, Inc.)
[8] Hurt M F and Waid C 1970 *SIAM J. Appl. Math.* **19** 547
[9] Marcus M and Minc H 1964 *A Survey of matrix theory and matrix inequalities* (Boston, Mass. : Allyn and Bacon)
[10] Salkin H M 1975 *Integer programming* (Reading, Mass. : Addison-Wesley)
[11] Sen S K 1978 *J. Indian Inst. Sci.* **61** 41
[12] Sen S K and Shamim A A 1978 *J. Indian Inst. Sci.* **60** 111
[13] Strum J E 1972 *Introduction to linear programming* (San Francisco : Holden-day)
[14] Vajda S 1974 *Theory of linear and nonlinear programming* (London : Longman)
[15] Vajda S 1975 *Problems in linear and nonlinear programming* (London : Charles Griffin)

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 35–42.

On duality in linear fractional programming

C R SESHAN
Department of Applied Mathematics, Indian Institute of Science, Bangalore 560 012

MS received 8 January 1979

Abstract. In this paper, a dual of a given linear fractional program is defined and the weak, direct and converse duality theorems are proved. Both the primal and the dual are linear fractional programs. This duality theory leads to necessary and sufficient conditions for the optimality of a given feasible solution. A unmerical example is presented to illustrate the theory in this connection. The equivalence of Charnes and Cooper dual and Dinkelbach's parametric dual of a linear fractional program is also established.

Keywords. Linear fractional programming; duality.

1. Introduction

In this paper a dual of a given linear fractional program is defined and this dual is also a linear fractional program. Kanti Swarup [11] has given a dual of a linear fractional program in which exists non-linearity in the constraints. Also, he did not prove the converse duality theorem. Kaska [7] has also given a dual of a linear fractional program which is constrained as the variable of the primal program. Chadha [2] has given a dual of a linear fractional program as a linear program which is nothing but the dual of the linear program obtained by Charnes and Cooper transformation of the original linear fractional program. Craven and Mond [4] have also given a dual of a linear fractional program such that both the primal and dual are linear fractional programs. Sharma and Swarup [10] have defined a dual of a linear fractional program in a different form but keeping the primal and dual as linear fractional programs.

Duals of nonlinear fractional programs have been proposed by Jagannathan [6], Bector [1], and Schaible [8,9]. The duals of Jagannathan and Schaible when applied to the linear fractional case give rise to the dual proposed by Chadha [2].

In the dual proposed by Sharma and Swarup [10], constant term does not appear in both the numerator and denominator of the objective function of the primal. This paper extends their definition to the general case where constant term is permitted to appear in the numerator and denominator of the objective function and the constraints of the dual are also generalised. This extension demands the revision of the proofs of the duality theorem. These proofs which make use of the results by Dinkelbach [5] are presented in this paper. This duality theory also

leads to a set of necessary and sufficient conditions for a feasible solution to be optimal and these conditions are extensions of Kuhn-Tucker necessary optimality conditions.

In the last section, it is proved that Charnes and Cooper dual of a linear fractional program as a linear program can be obtained independently by making use of the results proved by Dinkelbach [5] where the given linear fractional programming problem is converted into a parametric linear programming problem.

2. Dual of a linear fractional program

Consider the primal linear fractional programming problem (P1)

$$\text{Maximise } f(x) = (c^t x + a)/(d^t x + \beta)$$

(P1) subject to $Ax \leqslant b$

$$x \geqslant 0. \tag{1}$$

where A is an $(m \times n)$ matrix, c, d, x are $(n \times 1)$ vectors, b is an $(m \times 1)$ vector, a, β are scalars and t denotes transpose. Let

$$S = \{x \in R^n \mid Ax \leqslant b, x \geqslant 0\}.$$

Assume that S is nonempty and bounded and that f is not constant on S. Further assume that

$$d^t x + \beta > 0 \text{ for every } x \text{ in } R^n, \; x \geqslant 0.$$

This can be achieved if all the components of d are nonnegative and $\beta > 0$.

Define the dual (D1) corresponding to the primal (P1) as follows :

Minimise $g(u, v) = (c^t u + a)/(d^t u + \beta)$

subject to

(D1) $$c \cdot d^t u - d \cdot c^t u - A^t v \leqslant ad - \beta c \tag{2}$$

$$a \cdot d^t u - \beta \cdot c^t u + b^t v \leqslant 0 \tag{3}$$

$$u \geqslant 0, \; v \geqslant 0, \; u \in R^n, \; v \in R^m.$$

Theorem 1 (*Weak duality*)

If x is any feasible solution of (P1) and (u, v) is any feasible solution of (D1), then

$$f(x) \leqslant g(u, v).$$

Proof: Multiplying (2) by x^t

$$c^t x \cdot d^t u - d^t x \cdot c^t u - x^t A^t v \leqslant a \cdot d^t x - \beta \cdot c^t x \tag{4}$$

Multiplying (1) by v^t and using (3)

$$a \cdot d^t u - \beta \cdot c^t u \leqslant - b^t v \leqslant - x^t A^t v \tag{5}$$

From (4) and (5)

$$c^t x \cdot d^t u - d^t x \cdot c^t u + a \cdot d^t u - \beta \cdot c^t u \leqslant a \cdot d^t x - \beta \cdot c^t x \tag{6}$$

i.e. $$(c^t x + a)(d^t u + \beta) \leqslant (c^t u + a)(d^t x + \beta).$$

Hence,

$$f(x) = \frac{c^t x + a}{d^t x + \beta} \leqslant \frac{c^t u + a}{d^t u + \beta} = g(u, v).$$

Corollary 1 : If x is any feasible solution of (P1) and (u, v) is any feasible solution of (D1) such that $f(x) = g(u, v)$, then x solves (P1) and (u, v) solves (D1). Proof is obvious.

Theorem 2 (*Direct duality*)

If $\bar{x}$ solves (P1), then there exists $(\bar{u}, \bar{v})$ which solves (D1) such that $f(\bar{x}) = g(\bar{u}, \bar{v})$.

Proof. Let $\lambda = (c^t \bar{x} + a)/(d^t \bar{x} + \beta)$.

Consider the linear programming problem (P2).

Maximise $(c^t x + a) - \lambda (d^t x + \beta)$

(P2) subject to $Ax \leqslant b, x \geqslant 0$.

Dinkelbach [5] has proved that $\bar{x}$ also solves (P2) and the optimal value of the objective function in (P2) is 0.

Consider the dual of (P2) and denote it by (D2)

Minimise $b^t v + a - \lambda_\beta$

(D2) subject to $A^t v \geqslant c - \lambda d$ (7)

$$v \geqslant 0,\ v \in R^m$$

Let $\bar{v}_1$ solve (D2). By duality theorem of linear programming

$$b^t \bar{v}_1 + a - \lambda\beta = 0. \tag{8}$$

Let $\bar{u} = \bar{x}$ and $\bar{v} = \bar{v}_1 (d^t \bar{x} + \beta)$

$$\begin{aligned} c \cdot d^t \bar{u} - d \cdot c^t \bar{u} - A^t \bar{v} &\leqslant c \cdot d^t \bar{x} - d \cdot c^t \bar{x} - (c - \lambda d)(d^t \bar{x} + \beta) \\ &= c \cdot d^t \bar{x} - d \cdot c^t \bar{x} - c(d^t \bar{x} + \beta) + d(c^t \bar{x} + a) \\ &= ad - \beta c. \end{aligned}$$

Multiplying (8) by $d^t \bar{x} + \beta$,

$$b^t \bar{v} + a(d^t \bar{x} + \beta) - \beta (c^t \bar{x} + a) = b^t \bar{v} + a \cdot d^t \bar{u} - \beta \cdot c^t \bar{u} = 0.$$

Hence $(\bar{u}, \bar{v})$ is a feasible solution of (D1).

$$f(\bar{x}) = \frac{c^t \bar{x} + a}{d^t \bar{x} + \beta} = \frac{c^t \bar{u} + a}{d^t \bar{u} + \beta} = g(\bar{u}, \bar{v}),$$

$(\bar{u}, \bar{v})$ solves (D1), because of corollary.

Theorem 3 (*converse duality*)

If $(\bar{u}, \bar{v})$ solves (D1) then there exists an $\bar{x}$ which solves (P1) such that $f(\bar{x}) = g(\bar{u}, \bar{v})$.

Proof : Let $\lambda = (c^t \bar{u} + a)/(d^t \bar{u} + \beta)$.

Consider the linear programming problem (P3)

Minimise $(c^t u + a) - \lambda (d^t u + \beta)$

(P3) subject to $c \cdot d^t u - d \cdot c^t u - A^t v \leqslant ad - \beta c$

$$a \cdot d^t u - \beta \cdot c^t u + b^t v \leqslant 0$$

$$u \geqslant 0, v \geqslant 0.$$

Dinkelbach [5] has proved that $(\bar{u}, \bar{v})$ also solves (P3) and the optimal value of the objective function in (P3) is 0.

Consider the dual of (P3) and denote it by (D3)

Maximise $(-ad + \beta c)^t y + a - \lambda\beta$

subject to

$$(\text{D3}) (- c \cdot d^t + d \cdot c^t)^t y + (- a d + \beta c) \mu \leqslant c - \lambda d \tag{9}$$

$$Ay - b\mu \leqslant 0 \tag{10}$$

$$y \geqslant 0, \ \mu \geqslant 0, \ y \in R^n, \ \mu \in R.$$

Let $(\bar{y}, \bar{\mu})$ solve (D3).

By the duality theorem of linear programming

$$- ad^t \bar{y} + \beta \cdot c^t \bar{y} + a - \lambda\beta = 0 \tag{11}$$

$\bar{\mu} \neq 0$. For if $\bar{\mu} = 0$, from (10) we get $A\bar{y} \leqslant 0$, $\bar{y} \geqslant 0$. Since S is nonempty, there exists an x on S such that $Ax \leqslant b$, $x \geqslant 0$. Then $A(x + t\bar{y}) \leqslant b$, $x + t\bar{y} \geqslant 0$ for every $t > 0$. Hence $x + t\bar{y}$ is in S for every $t > 0$ which is a contradiction if $\bar{y} \neq 0$ since S is bounded.

If both $\bar{\mu} = 0$ and $\bar{y} = 0$, then from (11), $a - \lambda\beta = 0$. From (9), we get $c \geqslant \lambda d$. Let x be any feasible solution of (P1). We have $c^t x \geqslant \lambda d^t x$ and $a = \lambda\beta$. Hence $(c^t x + a) \geqslant \lambda (d^t x + \beta)$, i.e., $f(x) \geqslant \lambda$. But by Theorem 1, $f(x) \leqslant g(\bar{u}, \bar{v}) = \lambda$. Therefore $f(x) = \lambda$ for every feasible solution x of (P1) which implies that f is a constant on S, a contradiction to our assumption.

Therefore $\bar{\mu} > 0$. Let $\bar{x} = \bar{y}/\bar{\mu}$. From (10), $A\bar{x} \leqslant b$. Also $\bar{x} \geqslant 0$. Hence $\bar{x}$ is a feasible solution of (P1).

Applying complementary slackness theorem of linear programming to (P3) and (D3) we get

$$c^t \bar{y} \, d^t \bar{u} - d^t \bar{y} \, c^t \bar{u} - \bar{y}^t A^t \bar{v} - a \cdot d^t \bar{y} + \beta \cdot c^t \bar{y} = 0 \tag{12}$$

$$a\bar{\mu} \, d^t \bar{u} - \beta\bar{\mu} \, c^t \bar{u} + \bar{\mu} \, b^t \bar{v} = 0 \tag{13}$$

$$- d^t \bar{u} \, c^t \bar{y} + c^t \bar{u} \, d^t \bar{y} - a\bar{\mu} \ d^t \bar{u} + \beta\bar{\mu} \ c^t \bar{u} - c^t \bar{u} + \lambda d^t \bar{u} = 0 \tag{14}$$

$$\bar{y}^t A^t \bar{v} - \bar{\mu} \, b^t \bar{v} = 0. \tag{15}$$

Adding (12), (13) and (15)

$$c^t \bar{y} \, d^t \bar{u} - d^t \bar{y} \, c^t \bar{u} - ad^t \bar{y} + \beta c^t \bar{y} + a\bar{\mu} \, d^t \bar{u} - \beta\bar{\mu} c^t \bar{u} = 0 \tag{16}$$

$$(c^t \bar{y} + a\bar{\mu}) \ (d^t \bar{u} + \beta) = (d^t \bar{y} + \beta\bar{\mu}) (c^t \bar{u} + a).$$

Hence

$$\frac{c^t \bar{y} + a\bar{\mu}}{d^t \bar{y} + \beta\bar{\mu}} = \frac{c^t \bar{u} + a}{d^t \bar{u} + \beta}.$$

Therefore

$$f(x) = \frac{c^t \bar{x} + a}{d^t \bar{x} + \beta} = \frac{c^t \bar{u} + a}{d^t \bar{u} + \beta} = g(\bar{u}, \bar{v}).$$

Hence $\bar{x}$ solves (P1) because of the corollary.

Note 1 : Converse duality theorem can be deduced directly from direct duality theorem. This can be seen by the following argument. $(\bar{u}, \bar{v})$ solves (D1). Since S is compact, (P1) has a finite optimal solution say $\bar{x}$. By direct duality theorem, there exists $(\bar{u}_1, \bar{v}_1)$ which solves (D1) such that $f(\bar{x}) = g(\bar{u}_1, \bar{v}_1)$. Hence $f(x) = g(\bar{u}_1, \bar{v}_1) = g(\bar{u}, \bar{v})$.

Note 2 : We have not used the assumption that S is bounded in proving either the weak duality theorem or direct duality theorem. Even in the case of converse duality theorem, we can replace that assumption by the following weaker assumption

$$Ay \leqslant 0, \quad y \geqslant 0 \text{ implies } y = 0.$$

3. Some remarks

Remark 1 : The problem (P1) is equivalent to the problem (Q1)

Maximise $f(x) = (c^t x + a\, x_{n+1})/(d^t x + \beta x_{n+1})$

subject to $Ax \leqslant b$

$$x_{n+1} \leqslant 1$$

$$-x_{n+1} \leqslant -1$$

$$x \geqslant 0, \; x_{n+1} \geqslant 0, \; x \in R^n, \; x_{n+1} \in R.$$

In this form the problem (P1) is in the same form as the (LFP) considered by Sharma and Swarup [10]. We can write the dual (E1) of (Q1) as per definition of dual by Sharma and Swarup as follows:

Minimise $(c^t u + a u_{n+1})/(d^t u + \beta u_{n+1})$

subject to

(E1) $$c \cdot d^t u - d \cdot c^t u - A^t v \leqslant (ad - \beta c)\, u_{n+1} \tag{17}$$

$$a \cdot d^t u - \beta \cdot c^t u - v_{m+1} + v_{m+2} \leqslant 0 \tag{18}$$

$$b^t v + v_{m+1} - v_{m+2} \leqslant 0 \tag{19}$$

$$u, v, u_{n+1}, v_{m+1}, v_{m+2} \geqslant 0, \; u \in R^n, \; v \in R_m, \; u_{n+1}, \; v_{m+1}, \; v_{m+2} \in R.$$

Any feasible solution (u, v) of (D1) gives rise to many feasible solutions $(u', v', u'_{n+1}, v'_{m+1}, v'_{m+2})$ of (E1) with $u'_{n+1} = a$, where $a > 0$ is any real number,

$u' = a \cdot u$, $v = a \cdot v$, and v_{m+1}, v_{m+2} are chosen suitably. Also the corresponding objective function values become all equal. Conversely any feasible solution $(u', v', u'_{n+1}, v'_{m+1}, v'_{m+2})$ of (E1) with $u'_{n+1} \neq 0$ gives rise to a feasible solution (u, v) of (D1) where $u = u'/u'_{n+1}$, $v = v'/u'_{n+1}$ with the same objective function value. Feasible solutions of (E1) with $u'_{n+1} = 0$ do not correspond to any feasible solution of (D1). Therefore our dual (D1) is not equivalent to dual (E1). There is a one to many correspondence between feasible solutions of (D1) and a subset of feasible solutions of (E1). Thus (D1) is a much simpler dual than (E1) for the problem (P1).

Remark 2. The dual (R1) of the dual (D1) is

Maximise $(c^t x + \alpha)/(d^t x + \beta)$

subject to

$$\text{(R1)} \quad c \cdot d^t z - c \cdot d^t x - d \cdot c^t z + d \cdot c^t x \leqslant (ad - \beta c)(\lambda - 1) \tag{20}$$

$$a \cdot d^t z - a \cdot d^t x - \beta \cdot c^t z + \beta \cdot c^t x \leqslant 0 \tag{21}$$

$$Az - \lambda b \leqslant 0 \tag{22}$$

$$x, z, \lambda \geqslant 0, \; x, \; z \in R^n, \; \lambda \in R.$$

Any feasible solution x of (P1) gives rise to a feasible solution of (R1) if we take $z = x$ and $\lambda = 1$. Further the objective function values are equal. But the converse is not true. Hence dual of (D1) is not equivalent to (P1).

Remark 3. The above duality theory leads to necessary and sufficient conditions for a feasible solution x of the primal to be optimal. From the proof of direct duality theorem and weak duality theorem, it is easy to see that a feasible solution x of (P1) is an optimal solution of (P1) if and only if there exists a $v \geqslant 0$, $v \in R^m$ such that

$$c \cdot d^t x - d \cdot c^t x - A^t v \leqslant ad - \beta c \tag{23}$$

$$a \cdot d^t x - \beta \cdot c^t x + b^t v \leqslant 0. \tag{24}$$

Of the above two conditions, condition (23) is the Kuhn-Tucker necessary optimality conditions for (P1).

4. Numerical example

Primal problem

Maximise $(3x_1 + 3x_2 + 2x_3 + 1)/(2x_1 + x_2 + x_3 + 1)$

subject to $2x_1 + 5x_2 + x_3 \leqslant 2$

$x_1 + 2x_2 + 3x_3 \leqslant 3$

$x_1, x_2, x_3 \geqslant 0.$

Solving, we get that $x_1 = 0$, $x_2 = 3/13$, $x_3 = 11/13$ is an optimal solution and the optimal value of the objective function is 44/27.

Dual problem

Minimise $(3u_1 + 3u_2 + 2u_3 + 1)/(2u_1 + u_2 + u_3 + 1)$

subject to

$$-3u_2 - u_3 - 2v_1 - v_2 \leqslant -1$$

$$3u_1 + u_3 - 5v_1 - 2v_2 \leqslant -2$$

$$u_1 - u_2 - v_1 - 3v_2 \leqslant -1$$

$$-u_1 - 2u_2 - u_3 + 2v_1 + 3v_2 \leqslant 0$$

$$u_1, u_2, u_3, v_1, v_2 \geqslant 0.$$

Solving this we get that $u_1 = 0$, $u_2 = 3/13$, $u_3 = 11/13$, $v_1 = 7/13$, $v_2 = 1/13$ is an optimal solution and the optimal value of the objective function is 44/27. Other optimal solutions are:

$u_1 = 0$, $u_2 = 0$, $u_3 = 17/10$, $v_1 = 7/10$, $v_2 = 1/10$ and

$u_1 = 0$, $u_2 = 17/37$, $u_3 = 0$, $v_1 = 14/37$, $v_2 = 2/37$.

These results are as per the expectations of theorems 2 and 3.

5. A note on Charnes and Cooper dual of a linear fractional program

Charnes and Cooper (1962) converted the linear fractional programming problem (P1)

Maximise $(c^t x + \alpha)/(d^t x + \beta)$

(P1) subject to $Ax \leqslant b, x \geqslant 0$

under the assumption $d^t x + \beta > 0$ for every feasible x, into the linear programming problem (P4)

Maximise $c^t y + \alpha t$

(P4) subject to $Ay - bt \leqslant 0$

$$d^t y + \beta t = 1$$

$$y \geqslant 0, \; t \geqslant 0$$

using the transformation

$$t = 1/(d^t x + \beta), \; y = xt.$$

The Charnes and Cooper dual of (P1) is given by (D4)

Minimise λ

(D4) subject to $A^t u + \lambda d \geqslant c$

$$-b^t u + \lambda\beta \geqslant \alpha$$

$$u \geqslant 0, \; \lambda \text{ unrestricted}$$

Dinkelbach (1967) converted (Pl) into the parametric linear programming problem (P_λ)

$$\text{Maximise } (c^t x + a) - \lambda (d^t x + \beta)$$

(P_λ) subject to $Ax \leqslant b,\ x \geqslant 0.$

He proved that $\bar{x}$ solves (P1) and λ is the optimal value of the objective function in (P1) if and only if $\bar{x}$ solves (P_λ) and $F(\lambda) = 0$, where $F(\lambda)$ is the maximum of the objective function of (P_λ). He also proved that $F(\lambda)$ is a monotonic decreasing function of λ. Therefore the problem (P1) can be viewed as the one in which we have to find a λ such that $F(\lambda) = 0$ (i.e., to minimise λ such that $F(\lambda) \leqslant 0$). Considering the dual of (P_λ) we get

$$F(\lambda) = \min \{b^t u + a - \lambda\beta : A^t u \geqslant c - \lambda d, u \geqslant 0\}.$$

Hence the dual of the linear fractional programming problem (P1) is

Minimise λ

subject to $A^t u \geqslant c - \lambda d$

$$b^t u + a - \lambda\beta \leqslant 0.$$

$$u \geqslant 0, \lambda \text{ unrestricted.}$$

Hence Charnes and Cooper dual of a linear fractional program and Dinkelbach's parametric dual of a linear fractional program are essentially the same.

Acknowledgement

The author wishes to express his gratitude to Dr V G Tikekar, for his encouragement, guidance and for the fruitful discussions that the author had with him.

References

[1] Bector C R 1973 *Z. Opns. Res.* **17** 183
[2] Chadha S S 1971 *Z. Angew. Math. Mech.* **51** 560
[3] Charnes A and Cooper W W 1962 *Nav. Res. Log. Q.* 9, 181
[4] Craven B D and Mond B 1973 *J. Math. Anal. Appl.* **42** 507
[5] Dinkelbach W 1967 *Manage. Sci.* **13** 492
[6] Jagannathan R 1973 *Z. Opns. Res.* **17** 1
[7] Kaska J 1969 *Econ. Mat. Obzor.* **5** 442
[8] Schaible S 1974 *Z. Opns. Res.* **17** 187
[9] Schaible S 1976 *Manage. Sci.* **22** 858
[10] Sharma I C and Swarup K 1972 *Z. Opns. Res.* **16** 91
[11] Swarup K 1968 *Unternehmensforchung* **12** 106

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 43–52.

On generalised thermoelastic wave propagation

D S CHANDRASEKHARAIAH
Department of Mathematics, Bangalore University, Bangalore 560 001

MS received 21 July 1979

Abstract. Generalised thermoelasticity theories are employed to study one-dimensional disturbances in a half-space due to a thermal impulse on the boundary. Short time approximation of solutions are deduced and the exact discontinuities in the mechanical and thermal fields are analysed using the Laplace transform technique.

Keywords. Generalised thermoelasticity; relaxation time; wave motion.

1. Introduction

The theory of thermoelasticity which takes account of the time needed for the acceleration of heat flow has aroused much interest in recent years. This theory is a generalisation of the coupled thermoelasticity theory [5] and the field equations of this theory have been obtained by several authors on different grounds. For example, Lord and Shulman [13] have obtained the equations by generalising the energy equation of the coupled theory, while Green and Lindsay [7] have obtained the equations by generalising both the constitutive equation and the energy equation. The theory developed in [13] involves one relaxation time of the thermoelastic process and that obtained in [7] involves two relaxation times. Several problems revealing interesting phenomena which characterise the generalised thermoelasticity theory have been considered [1]–[4], [6], [10]–[21].

In this paper we investigate one-dimensional disturbances in a thermoelastic half-space with plane boundary by employing the equations obtained in [7] and [13]. We suppose that the half-space occupies the region $x \geqslant 0$, and that initially it is at rest in its undeformed state and is maintained at a uniform (reference) temperature and zero temperature-velocity. At time $t = 0$, the boundary $x = 0$ is subjected to a distribution of temperature given by the Dirac-delta function and is held at zero strain. We determine the mechanical and thermal fields due to the disturbances produced. We solve the basic equations by the Laplace transform technique and under short time approximation we find in the contexts of both the theories that the disturbances consist of two waves propagating with finite speeds. One of these waves is the elastic wave influenced by the thermal field and the other is the thermal wave, and the former wave follows the latter wave. Further, the elastic wave of the Lord–Shulman theory is faster than that of the Green–Lindsay

theory. In the contexts of both the theories, the temperature, strain and stress all experience discontinuities at each of the wavefronts. We obtain and analyse the exact magnitudes of these discontinuities in the contexts of both the theories and compare the results. The corresponding results of the coupled theory are also deduced.

2. Analysis on the basis of the Lord–Shulman Theory

Under the assumptions made in section 1, the half-space experiences one-dimensional deformation parallel to the x-axis. In the context of the linearised Lord-Shulman theory, the equations governing the displacement U, strain E, stress Σ and the temperature T (above the reference temperature T_0) are given by [13]

$$E = \frac{\partial U}{\partial x} \tag{1}$$

$$\Sigma = (\lambda + 2\mu) E - \nu T \tag{2}$$

$$(\lambda + 2\mu) \frac{\partial^2 U}{\partial x^2} - \nu \frac{\partial T}{\partial x} = \rho \frac{\partial^2 U}{\partial t^2} \tag{3}$$

$$k \frac{\partial^2 T}{\partial x^2} = \left(1 + \tau \frac{\partial}{\partial t}\right) \left(C \frac{\partial T}{\partial t} + \nu T_0 \frac{\partial E}{\partial t}\right). \tag{4}$$

In these equations λ and μ are the Lamé constants, $\nu = (3\lambda + 2\mu)\gamma$, γ being the coefficient of linear thermal expansion, C is the specific heat at constant volume, ρ is the mass density and τ is the relaxation time. It is assumed that there are no body forces and there are no heat sources.

Putting $C_1^2 = (\lambda + 2\mu)/\rho$

and applying the transformations

$$\left.\begin{aligned} x &\to \frac{k}{CC_1} x, \qquad & t &\to \frac{k}{CC_1^2} t \\ U &\to \frac{k}{CC_1} U, \qquad & T &\to \frac{\rho C_1^2}{\nu} T \\ E &\to E, \qquad & \Sigma &\to (\lambda + 2\mu)\Sigma \end{aligned}\right\} \tag{5}$$

to the equations (1)–(4), we obtain the following equations which are in dimensionless form.

$$E = \frac{\partial U}{\partial x} \tag{6}$$

$$\Sigma = E - T \tag{7}$$

$$\frac{\partial^2 U}{\partial x^2} - \frac{\partial T}{\partial x} = \frac{\partial^2 U}{\partial t^2} \tag{8}$$

$$\frac{\partial^2 T}{\partial x^2} = \left(1 + \beta \frac{\partial}{\partial t}\right) \left(\frac{\partial T}{\partial t} + \epsilon \frac{\partial E}{\partial t}\right). \tag{9}$$

Here we have put

$$\epsilon = \frac{\nu^2 T_0}{\rho C C_1^2}, \quad \beta = \frac{C C_1^2}{k}\tau. \tag{10}$$

Under the assumptions made in section 1, the initial and boundary conditions are given by

$$T(x,0) = \frac{\partial T}{\partial t}(x,0) = U(x,0) = \frac{\partial U}{\partial t}(x,0) = 0 \tag{11}$$

$$E(o,t) = 0, \quad T(o,t) = \delta(t) \tag{12}$$

where $\delta(t)$ is the Dirac-delta function.

We now solve the equations (6)-(9) under the conditions (11) and (12) by the Laplace transform method. Taking the Laplace transform of the equation (6), (8) and (9) under the conditions (11) and solving the resulting equations simultaneously under the conditions (12) we arrive at the following solutions for E and T.

$$\bar{T}(x,p) = \frac{1}{m_1^2 - m_2^2}\left[\{m_1^2 - p^2 - \epsilon p(1+\beta p)\}\right] e^{-m_1 x}$$
$$\{m_2^2 - p^2 - \epsilon p(1+\beta p)\}\, e^{-m_2 x}] \tag{13}$$

$$\bar{E}(x,p) = \frac{p(1+\beta p)}{m_1^2 - m_2^2}\left[e^{-m_2 x} - e^{-m_1 x}\right]. \tag{14}$$

Here m_1 and m_2 satisfy the equations

$$\left.\begin{aligned} m_1^2 + m_2^2 &= p[1 + \epsilon + p(1+\beta+\epsilon\beta)] \\ m_1^2 - m_2^2 &= p[\{1+\epsilon+p(1+\beta+\epsilon\beta)\}^2 - 4p(1+\beta p)]^{1/2} \end{aligned}\right\} \tag{15}$$

and $\bar{f}(x,p)$ is the Laplace transform of $f(x,t)$.

The solutions (13) and (14) are not readily invertible. Since the effects of relaxation time on thermoelastic interactions are short lived [7], we restrict ourselves to the short time approximation of the solution. This approximation corresponds to $p \to \infty$.

In the limit $p \to \infty$, we obtain the following expressions from (15).

$$m_{1,2} = \frac{p}{v_{1,2}} + q_{1,2} + \frac{L_{1,2}}{p} + O\left(\frac{1}{p^2}\right) \tag{16}$$

$$(m_1^2 - m_2^2)^{-1} = p^{-3}\Gamma^{-3/2}\left[p\Gamma - M - \frac{N}{p} + O\left(\frac{1}{p^2}\right)\right] \tag{17}$$

where

$$v_{1,2} = \sqrt{2}\,[1 + \beta + \epsilon\beta \pm \Gamma^{1/2}]^{-1/2} \tag{18}$$

$$\left.\begin{aligned} q_{1,2} &= \tfrac{1}{4} v_{1,2}[1 + \epsilon \pm M\Gamma^{-1/2}] \\ L_{1,2} &= \pm \tfrac{1}{4} v_{1,2} N\Gamma^{-1/2} \\ \Gamma &= (1 - \beta + \epsilon\beta)^2 + 4\epsilon\beta^2 \\ M &= \beta(1+\epsilon)^2 + \epsilon - 1 \\ N &= \tfrac{1}{2}[(1+\epsilon)^2 - 3\Gamma^{-1}M^2]. \end{aligned}\right\} \tag{19}$$

The expressions for m_1 and m_2 given by (16) show that the solutions (13) and (14) consist of two waves propagating with speeds v_1 and v_2. From (18) we find that $v_1 \to 1$ and $v_2 \to \infty$, as $\beta \to 0$. Since $\beta = 0$ corresponds to the case of the coupled theory which predicts an infinite speed for heat propagation, we may conclude that the wave which propagates with speed v_2 is the thermal wave. The other wave is obviously the elastic wave influenced by the thermal field. Since $v_1 < v_2$ by (18), we may infer that the elastic wave follows the thermal wave. As a consequence of this, the points of the half-space for which $x > tv_2$ do not experience any disturbance.

Inverting the solutions (13) and (14) with the help of (17), we obtain

$$\begin{aligned} T(x,t) = {} & \tfrac{1}{2}\Gamma^{-1/2} e^{-q_1 x}\Big[\{1-\epsilon+M\Gamma^{-1}(1-\beta+\epsilon\beta)\} \\ & + \tfrac{1}{2}\Gamma^{-1}\{\Gamma^{1/2}(1+\epsilon)^2 + M^2\Gamma^{-1/2} - 2M(1-\epsilon) \\ & + 2N(1-\beta+\epsilon\beta-\Gamma^{1/2})\}\left(t-\frac{x}{v_1}\right)\Big] H\left(t-\frac{x}{v_1}\right) \\ & - \tfrac{1}{2}\Gamma^{-1/2} e^{-q_2 x}\Big[\{1-\epsilon+M\Gamma^{-1}(1-\beta+\epsilon\beta)\} \\ & - \tfrac{1}{2}\Gamma^{-1}\{\Gamma^{1/2}(1+\epsilon)^2 + M^2\Gamma^{-1/2} + 2M(1-\epsilon) \\ & - 2N(1-\beta+\epsilon\beta+\Gamma^{1/2})\}\left(t-\frac{x}{v_2}\right)\Big] H\left(t-\frac{x}{v_2}\right). \end{aligned} \tag{20}$$

$$\begin{aligned} E(x,t) = {} & \Gamma^{-3/2} e^{-q_1 x}\Big[(\Gamma-\beta M) - (M+\beta N)\left(t-\frac{x}{v_1}\right)\Big] H\left(t-\frac{x}{v_1}\right) \\ & - \Gamma^{-3/2} e^{-q_2 x}\Big[(\Gamma-\beta M) - (M+\beta N)\left(t-\frac{x}{v_2}\right)\Big] H\left(t-\frac{x}{v_2}\right) \end{aligned} \tag{21}$$

where $H(x)$ is the Heaviside unit function.

Equations (7), (20) and (21) now yield

$$\begin{aligned} \Sigma(x,t) = {} & \tfrac{1}{2}\Gamma^{-1/2} e^{-q_1 x}\Big[\{1+\epsilon-M\Gamma^{-1}(1+\beta+\epsilon\beta)\} \\ & - \tfrac{1}{2}\Gamma^{-1}(1+\epsilon)^2\Gamma^{1/2} + M^2\Gamma^{-1/2} + 2M(1+\epsilon) \\ & + 2N(1+\beta+\epsilon\beta-\Gamma^{1/2})\}\left(t-\frac{x}{v_1}\right)\Big] H\left(t-\frac{x}{v_1}\right) \\ & - \tfrac{1}{2}\Gamma^{-1/2} e^{-q_2 x}\Big[\{(1+\epsilon-M\Gamma^{-1}(1+\beta+\epsilon\beta)\} \\ & + \tfrac{1}{2}\Gamma^{-1}\{(1+\epsilon)^2\Gamma^{1/2} + M^2\Gamma^{-1/2} - 2M(1+\epsilon) \\ & - 2N(1+\beta+\epsilon\beta+\Gamma^{1/2})\}\left(t-\frac{x}{v_2}\right)\Big] H\left(t-\frac{x}{v_2}\right). \end{aligned} \tag{22}$$

In the solutions (20)–(22), the first term represents the contribution of the elastic wave in the vicinity of its wavefront ($x = v_1 t$) and the second term represents the

contribution of the thermal wave in the vicinity of its wavefront ($x = v_2 t$). We readily see that the temperature, strain and stress all experience discontinuities at each of the wavefronts. The magnitudes of these discontinuities are given by

$$[T^+ - T^-]_{x=tv_{1,2}} = \pm \tfrac{1}{2} \Gamma^{-1/2} [1 - \epsilon + M\Gamma^{-1} (1 - \beta + \epsilon\beta)] \, e^{-a_{1,2}x} \tag{23}$$

$$[E^+ - E^-]_{x=tv_{1,2}} = \pm \Gamma^{-1/2} [1 - \beta M \Gamma^{-1}] \, e^{-a_{1,2}x} \tag{24}$$

$$[\Sigma^+ - \Sigma^-]_{x=tv_{1,2}} = \pm \tfrac{1}{2} \Gamma^{-1/2} [1 + \epsilon - M\Gamma^{-1} (1 + \beta + \epsilon\beta)] \, e^{-a_{1,2}x} \tag{25}$$

In view of the Tauberian theorem for hyperbolic equations [9] we note that although the solutions (20)–(22) are valid for relatively short times, the magnitudes of discontinuities given by (23)–(25) are exact.

We consider two particular cases.

Case (i) : $\beta = 0$. In this case the Lord–Shulman theory reduces to the coupled theory. Equations (18) now yield $v_1 = 1$ and $v_2 \to \infty$. We verify from equations (23)-(25) that the thermal field is continuous at the elastic wavefront and the magnitudes of discontinuities of strain and stress at this wavefront are given by

$$[E^+ - E^-]_{x=t} = [\Sigma^+ - \Sigma^-]_{x=t} = \exp(-\tfrac{1}{2} \epsilon x). \tag{26}$$

Thus, in the coupled theory four of the six discontinuities given by (23)–(25) disappear. Further, from (26) it follows that the strain and stress effects occur instantaneously throughout the body. This contrasts with the general case in which no effect is observed for $x > tv_2$.

Case (ii) : $\epsilon = 0$. In this case the coupling between the thermal field and the elastic field ceases. Equation (18) now yield $v_1 = 1$ and $v_2 = \beta^{-\frac{1}{2}}$. We verify from (23)–(25) that the thermal field is continuous at both the wavefronts. The magnitudes of discontinuities of stress and strain are given by

$$[E^+ - E^-]_{x=t} = [\Sigma^+ - \Sigma^-]_{x=t} = (1 - \beta)^{-2} \tag{27}$$

$$[E^+ - E^-]_{x=t/(\beta)^{\frac{1}{2}}} = [\Sigma^+ - \Sigma^-]_{x=t/(\beta)^{\frac{1}{2}}} = -(1 - \beta)^{-2} \exp\left(-\frac{x}{2\sqrt{\beta}}\right) \tag{28}$$

Thus in this case two of the six discontinuities disappear. The strain and stress have discontinuities of equal magnitude at both the wavefronts. The discontinuities are uniform at the elastic wavefront and decay exponentially with distance at the thermal wavefront.

3. Analysis on the basis of the Green-Lindsay theory

In the context of the linearised Green-Lindsay theory, the equations governing one-dimensional deformation parallel to the x-axis are [7], [8]

$$E = \frac{\partial U}{\partial x} \tag{29}$$

$$\Sigma = (\lambda + 2\mu) E - \nu \left(1 + a \frac{\partial}{\partial t}\right) T \tag{30}$$

$$(\lambda + 2\mu)\frac{\partial^2 U}{\partial x^2} - \nu\left(1 + a\frac{\partial}{\partial t}\right)\frac{\partial T}{\partial x} = \rho\frac{\partial^2 U}{\partial t^2} \tag{31}$$

$$k\frac{\partial^2 T}{\partial x^2} = C\left(1 + a^*\frac{\partial}{\partial t}\right)\frac{\partial T}{\partial t} + \nu T_0\frac{\partial E}{\partial t} \tag{32}$$

Here a and a^* are relaxation times.

Under the transformations (5), equations (29)–(32) reduce to the following dimensionless form.

$$E = \frac{\partial U}{\partial x} \tag{33}$$

$$\Sigma = E - \left(1 + \delta\frac{\partial}{\partial t}\right)T \tag{34}$$

$$\frac{\partial^2 U}{\partial x^2} - \left(1 + \delta\frac{\partial}{\partial t}\right)\frac{\partial T}{\partial x} = \frac{\partial^2 U}{\partial t^2} \tag{35}$$

$$\frac{\partial^2 T}{\partial x^2} = \left(1 + \eta\frac{\partial}{\partial t}\right)\frac{\partial T}{\partial t} + \epsilon\frac{\partial E}{\partial t} \tag{36}$$

Here we have put

$$\delta = \frac{CC_1^2}{k}a, \quad \eta = \frac{CC_1^2}{k}a^*. \tag{37}$$

It has been found [8] that a and a^* satisfy the inequalities $a \geqslant a^* \geqslant 0$. Accordingly, we have

$$\delta \geqslant \eta \geqslant 0. \tag{38}$$

The initial and boundary conditions in the present case are also as given by (11) and (12).

Taking the Laplace transform of the equations (29), (31) and (32) under the conditions (11) and solving the resulting equations simultaneously under the conditions (12), we obtain the following solutions for T and E.

$$\bar{T}(x, p) = \frac{1}{n_1^2 - n_2^2}\left[\{n_1^2 - p^2 - \epsilon p(1 + \delta p)\}\, e^{-n_1 x}\right.$$
$$\left. - \{n_2^2 - p^2 - \epsilon p(1 + \delta p)\}\, e^{-n_2 x}\right] \tag{39}$$

$$\bar{E}(x, p) = \frac{p(1 + \eta p)(1 + \delta p)}{n_1^2 - n_2^2}\left[e^{-n_1 x} - e^{-n_2 x}\right] \tag{40}$$

where

$$\left.\begin{aligned} n_1^2 + n_2^2 &= p\left[1 + \epsilon + p(1 + \eta + \epsilon\delta)\right] \\ n_1^2 - n_2^2 &= p\left[\{1 + \epsilon + p(1 + \eta + \epsilon\delta)\}^2 - p(1 + \eta p)\right]^{1/2} \end{aligned}\right\} \tag{41}$$

For small values of t (i.e. for $p \to \infty$), the expressions (41) yield

$$n_{1,2} = \frac{p}{V_{1,2}} + r_{1,2} + \frac{Q_{1,2}}{p} + O\left(\frac{1}{p^2}\right) \tag{42}$$

$$(n_1^2 - n_2^2)^{-1} = p^{-3}\,\Delta^{-3/2}\left[p\Delta - R - \frac{S}{p} + O\left(\frac{1}{p^2}\right)\right] \tag{43}$$

where

$$V_{1,2} = \sqrt{2}\,[1 + \eta + \epsilon\delta \pm \triangle^{1/2}]^{-1/2} \tag{44}$$

$$\left.\begin{aligned} r_{1,2} &= \tfrac{1}{4}\, V_{1,2}\,[1 + \epsilon \pm R\triangle^{-1/2}] \\ Q_{1,2} &= \pm\tfrac{1}{4}\, V_{1,2}\, S\triangle^{-1/2} \\ \triangle &= (1 - \eta + \epsilon\delta)^2 + 4\epsilon\delta\eta \\ R &= (1 + \epsilon)(\eta + \epsilon\delta) + \epsilon - 1 \\ S &= \tfrac{1}{2}[(1 + \epsilon)^2 - 3R^2\,\triangle^{-1}]. \end{aligned}\right\} \tag{45}$$

The form of expressions (42) show that the solutions (39) and (40) consist of two waves propagating with speeds V_1 and V_2 given by (44). We verify that the wave which propagates with speed V_1 is the elastic wave influenced by the thermal field and the wave which propagates with speed V_2 is the thermal wave. As in the Lord-Shulman theory, the elastic wave follows the thermal wave.

From equations (34) and (44) we see that if $\delta \neq 0$, the stress depends on the temperature-velocity and if $\eta \neq 0$, the heat propagates with a finite speed. Since $\eta \neq 0$ implies $\delta \neq 0$, because of (38), it follows that the heat cannot propagate with a finite speed unless the stress depends on the temperature-velocity. This is not the case with the Lord-Shulman theory wherein the heat can propagate with a finite speed although the stress there is independent of the temperature-velocity.

In the particular case when $\delta = \eta$, the speeds of propagation are governed by the expressions

$$\left.\begin{aligned} V_{1,2} &= \sqrt{2}\,[1 + \eta + \epsilon\eta \pm \triangle^{1/2}]^{-1/2} \\ \triangle &= (1 - \eta + \epsilon\eta)^2 + 4\epsilon\eta^2. \end{aligned}\right\} \tag{46}$$

Comparing these expressions with (18) and (19) and identifying η with β, we find that $V_{1,2} = v_{1,2}$. Since we have $\delta \geqslant \eta \geqslant 0$, it follows that $V_1 \leqslant v_1$. Accordingly, in general, the elastic wave in the Green-Lindsay theory is slower than its counterpart in the Lord-Shulman theory.

Inverting the solutions (39) and (40) with the help of (43) we obtain

$$\begin{aligned} T(x,t) = {} & \tfrac{1}{2}\triangle^{-1/2}\, e^{-r_1 x}\Big[\{1 - \epsilon + R\triangle^{-1}(1 - \eta + \epsilon\delta)\} \\ & + \tfrac{1}{2}\triangle^{-1}\{(1 + \epsilon)^2\,\triangle^{1/2} - 2R(1 - \epsilon) + \triangle^{-1/2}\,R^2 + 2S(1 - \eta + \epsilon\delta \\ & - \triangle^{1/2})\}\Big(t - \frac{x}{V_1}\Big)\Big]\, H\Big(t - \frac{x}{V_1}\Big) - \tfrac{1}{2}\,\triangle^{-1/2}\, e^{-r_2 x} \\ & \times\Big[\{1 - \epsilon + R\triangle^{-1}(1 - \eta + \epsilon\delta)\} - \tfrac{1}{2}\triangle^{-1}\{(1 + \epsilon)^2\,\triangle^{1/2} \\ & + 2R(1 - \epsilon) + R^2\,\triangle^{-1/2} - 2S(1 - \eta + \epsilon\delta + \triangle^{1/2})\}\Big(t - \frac{x}{V_2}\Big)\Big] \\ & \times H\Big(t - \frac{x}{V_2}\Big) \end{aligned} \tag{47}$$

$$E(x,t) = \triangle^{-3/2} e^{-r_1 x}\Big[\{\triangle - (\delta+\eta)R - \delta\eta S\} - \{R + S(\delta+\eta)\} \times \Big(t - \frac{x}{V_1}\Big)\Big] H\Big(t - \frac{x}{V_1}\Big) - \triangle^{-3/2} e^{-r_2 x} \times \Big[\{\triangle - (\delta+\eta)R - \delta\eta S\} - \{R + S(\delta+\eta)\}\Big(t - \frac{x}{V_2}\Big)\Big] \times H\Big(t - \frac{x}{V_2}\Big) \tag{48}$$

Taking the Laplace transform of the equation (34), substituting for $\bar{T}(x,p)$ and $\bar{E}(x,p)$ from (39) and (40) and then inverting the resulting expression with the aid of (43) we obtain

$$\begin{aligned}\Sigma(x,t) = {} & \tfrac{1}{2}\triangle^{-1/2} e^{-r_1 x}\Big[(1+\epsilon)\{1 - \delta R\triangle^{-1} - \tfrac{1}{2}(1+\epsilon)\delta\triangle^{-1/2}\} \\ & - R\triangle^{-1/2}(1 + \tfrac{1}{2}\delta R\triangle^{-1}) - \triangle^{-1}(R+\delta S)(1+\eta+\epsilon\delta - \triangle^{1/2}) \\ & - \{\tfrac{1}{2}(1 - \delta R\triangle^{-1})(\triangle^{-1/2}(1+\epsilon)^2 + 3R^2\triangle^{-3/2}) + \triangle^{-1}(R+\delta S) \\ & \times (1+\epsilon - R\triangle^{-1/2}) + R\triangle^{-1}(1+\eta+\epsilon\delta - \triangle^{1/2})\} \\ & \times \Big(t - \frac{x}{V_1}\Big)\Big] H\Big(t - \frac{x}{V_1}\Big) - \tfrac{1}{2}\triangle^{-1/2} e^{-r_2 x}\Big[(1+\epsilon)\{1 - \delta R\triangle^{-1} \\ & + \tfrac{1}{2}(1+\epsilon)\delta\triangle^{-1/2}\} + R\triangle^{-1/2}(1 + \tfrac{1}{2}\delta R\triangle^{-1}) - \triangle^{-1}(R+\delta S) \\ & \times (1+\eta+\epsilon\delta + \triangle^{1/2}) + \{\tfrac{1}{2}(1 - \delta R\triangle^{-1})(\triangle^{-1/2}(1+\epsilon)^2 \\ & + 3R^2\triangle^{-3/2}) - \triangle^{-1}(R+\delta S)(1+\epsilon + R\triangle^{-1/2}) - R\triangle^{-1} \\ & \times (1+\eta+\epsilon\delta + \triangle^{1/2})\}\Big(t - \frac{x}{V_2}\Big)\Big] H\Big(t - \frac{x}{V_2}\Big). \end{aligned} \tag{49}$$

In (47)–(49) the first term represents the contribution of the elastic wave in the vicinity of the wavefront $x = V_1 t$ and the second term represents the contribution of the thermal wave at the vicinity of the wavefront $x = V_2 t$.

We readily see from (47)–(49) that the temperature, strain and stress all experience discontinuities at both the wavefronts, as in the Lord-Shulman theory. The magnitudes of these discontinuities are given by

$$[T^+ - T^-]_{x=tV_{1,2}} = \pm\tfrac{1}{2}\triangle^{-1/2}[1 - \epsilon + R\triangle^{-1}(1 - \eta + \epsilon\delta)]\, e^{-r_{1,2}x} \tag{50}$$

$$[E^+ - E^-]_{x=tV_{1,2}} = \pm\triangle^{-3/2}[\triangle - R(\delta+\eta) - \delta\eta S] e^{-r_{1,2}x} \tag{51}$$

$$\begin{aligned}[\Sigma^+ - \Sigma^-]_{x=tV_{1,2}} = {} & \pm\tfrac{1}{2}\triangle^{-1/2}[(1+\epsilon)\{1 - \delta R\triangle^{-1} \mp \tfrac{1}{2}(1+\epsilon)\delta\triangle^{-1/2}\} \\ & \mp R\triangle^{-1/2}(1 + \tfrac{1}{2}\delta R\triangle^{-1}) - \triangle^{-1}(R+S\delta)(1+\eta+\epsilon\delta \\ & \mp \triangle^{1/2})]\, e^{-r_{1,2}x}. \end{aligned} \tag{52}$$

As in section 2 we note that although the solutions (47)–(49) are just short time approximations, the magnitudes of discontinuities given by (50)–(52) are exact.

We consider three particular cases.

Case (i) : $\eta = 0$. In this case we have $V_1 = (1 + \epsilon\delta)^{-1/2}$ and $V_2 \to \infty$. Hence the heat propagates with infinite speed. The expressions (50)–(52) now yield the following discontinuities at the elastic wavefront.

$$[T^+ - T^-]_{x=tV_1} = \frac{\epsilon\delta}{(1+\epsilon\delta)^2} e^{-r_1 x}$$

$$[E^+ - E^-]_{x=tV_1} = \frac{1 + \delta(1 + \epsilon - \epsilon\delta)}{(1+\epsilon\delta)^3} e^{-r_1 x}$$

$$[\Sigma^+ - \Sigma^-]_{x=tV_1} = \tfrac{1}{2}(1+\epsilon\delta)^{-2}\left[2 - \frac{\delta}{2}\{(1+\epsilon)^2 + R^2 \triangle^{-1} + 2(1+\epsilon) R \triangle^{-1/2}\}\right] e^{-r_1 x}$$

with $$r_1 = \frac{\epsilon V_1}{2}\left(1 + \frac{\delta}{1+\epsilon\delta}\right).$$

Thus when the heat propagates with infinite speed, unlike in the Lord-Shulman theory, the temperature, strain and stress all experience discontinuity at the elastic wavefront and the magnitude of discontinuity of stress is not equal to that of strain. Obviously, this deviation is because of the dependence of stress on the temperature-velocity.

In the limiting case when $\delta \to 0$, the thermal field is continuous at the elastic wavefront and the discontinuities of stress and strain reduce to those of the coupled theory, *viz.*, the expressions given by (26).

Case (ii) : $\delta = \eta \neq 0$. In this case the expressions (47) and (50) associated with the thermal field reduce to the corresponding expressions obtained in section 2. The expressions for stress and strain, however, differ from those obtained there.

Case (iii) : $\epsilon = 0$. In this case we obtain from equations (44), $V_1 = 1$ and $V_2 = \eta^{-1/2}$. We verify that the results obtained in case (iii) of section 2 hold here also.

References

[1] Agrawal V K 1979 *Acta Mech.* **31** 185
[2] Agrawal V K 1978 *J. Elast.* **8** 171
[3] Boschi E 1972 *Lett. Nuovo Cimento* **5** 192
[4] Boschi E 1974 *Int. J. Eng. Sci.* **12** 433
[5] Chadwick P 1960 *Progress in Solid Mechanics* (ed. I N Sneddon and R Hill) (Amsterdam: N H Publ. Co.) p 265
[6] Fox N 1969 *Int. J. Eng. Sci.* **7** 437
[7] Green A E and Lindsay K A 1972 *J. Elast.* **2** 1
[8] Green A E 1972 *Mathematika* **19** 69
[9] Knopoff L and Gilbert F 1959 *J. Acoust. Soc. Am.* **31** 1161
[10] Kolyano Yu M and Semerak F V 1972 *Sov. Mat. Sci.* **8** 79
[11] Kolyano Yu M and Semerak F V 1973 *Strength Mater.* **5** 234
[12] Kotenko N V and Lenyuk M P 1975 *Sov. Appl. Mech.* **10** 147
[13] Lord H W and Shulman Y 1967 *J. Mech. Phys. Solids* **15** 299

[14] McCarthy M F 1972 *Int. J. Eng. Sci.* **10** 593
[15] Norwood F R and Warren W E 1969 *Q. J. Mech. appl. Math.* **22** 283
[16] Nayfeh A H and Nemat-Nasser S 1971 *Acta Mech.* **12** 53
[17] Nayfeh A H and Nemat-Nasser S 1972 *Trans. ASME. Ser. E.* **39** 108
[18] Popov E B 1967 *PMM* **31** 349
[19] Puri P 1973 *Int. J. Eng. Sci.* **11** 735
[20] Wadhawan M C 1972 *Indian J. Pure Appl. Math.* **3** 1010
[21] Wadhawan M C 1973 *Pure Appl. Geophys.* **102** 37

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 53–60.

On weak discontinuities through thermally conducting and dissociating gases

RAMA SHANKAR and SUNIL KUMAR JAIN
Department of Mathematics, Indian Institute of Technology, New Delhi 110 029

MS received 10 July 1978; revised 16 July 1979

Abstract. Using singular surface theory, the phenomena associated with the uniform and nonuniform propagation of weak discontinuities through thermally conducting and dissociating gases is studied. The basic differential equations governing the growth and decay of weak discontinuities are derived and solved completely. The criteria for decay or blow up of these discontinuities is obtained. It turns out that the thermal conduction and dissociation allow the existence of a singular surface carrying a weak discontinuity which grows into a shock and the role of dissociation and thermal conduction is to cause rapid damping in the formation of this shock.

Keywords. Singular surface; thermally conducting; dissociating gases; shock waves.

1. Introduction

It was Thomas [17], [18], who first introduced singular surface theory to study the propagation of weak discontinuities through uniform media. Many applications of his method followed (Kaul [7]; Nariboli [12]; Shankar [15]; Srinivasan [16]). Ludwig [10] and Duff [5] further generalized and developed this technique. Lighthill [9] studied the anisotropic wave propagation and provided the asymptotic feature of the solution. Bazer [1] successfully used the mathematical theory of geometric optics (ray theory), given by Luneberg [11], to investigate the propagation of weak discontinuities. Nariboli [13], [14] combined the singular surface theory with the ray theory to study the propagation of weak discontinuities in nonlinear anisotropic media and was able to integrate the growth equation. Following Nariboli [14], Upadhyay [20] obtained the growth equation for sonic discontinuities propagating through thermally conducting gases. Recently Elcrat [6] studied the nonuniform propagation of sonic discontinuities in an unsteady flow of a perfect gas.

The present paper extends the singular surface theory to study the phenomena associated with uniform as well as nonuniform propagation of weak discontinuities through thermally conducting and dissociating gases. The plan of the paper is as follows :

The equations of motion governing the three dimensional, unsteady, thermally conducting and dissociating gas flow are given in § 2. Using compatibility condi-

tions in § 3, it comes out that the weak discontinuities propagate with the isothermal frozen speed of sound. Further, basic differential equations for growth and decay of sonic discontinuities for uniform medium are derived and solved completely in § 4 while nonuniform case has been discussed in § 5. In order to integrate the resultant transport equation, we transform it to an equation along the bicharacteristic curve in the characteristic manifold to find the criterion for decay or blow up of sonic discontinuities. Some particular cases are discussed to predict the physical condition more clearly.

2. The basic equations

The fundamental equations of motion for unsteady, three dimensional, thermally conducting and dissociating gas flow are given by

$$\frac{\partial \rho}{\partial t} + v_i \rho_{,i} + \rho v_{i,i} = 0 \tag{1}$$

$$\rho \frac{\partial v_i}{\partial t} + \rho v_k v_{i,k} + p_{,i} = 0 \tag{2}$$

$$\rho \frac{\partial h}{\partial t} + \rho v_i h_{,i} - \frac{\partial p}{\partial t} - v_i p_{,i} = K\theta_{,ii} \tag{3}$$

$$p = \rho R\theta (1 + q) \tag{4}$$

$$h = R\theta (4 + q) + Dq \tag{5}$$

and

$$\frac{\partial q}{\partial t} + v_i q_{,i} = C\rho\theta^{-n} \left\{ (1 - q)\, e^{-D/R\theta} - \frac{\rho}{\rho_D} q^2 \right\} \tag{6}$$

where t denotes the time, $\partial/\partial t$ the partial differentiation with respect to t, ρ the density, p the pressure, θ the temperature, v_i the components of flow velocity, h the enthalpy, K the thermal conductivity, q the mass fraction variable, R the universal gas constant, D the dissociation energy per unit mass, ρ_D the characteristic density and C, a constant with the dimension of time. The repeated index implies the summation unless otherwise stated and a comma followed by an index denotes partial differentiation with respect to space variable.

From the above set of equations we deduce

$$\frac{\partial \rho}{\partial t} - \rho v_k \frac{\partial v_k}{\partial t} - \rho v_{k,i} v_k v_i + c_f^2 \rho v_{k,k} + c\rho\theta^{-n} \left\{ \rho D \frac{(1 + q)}{3} - \rho R\theta \right\}$$

$$\times \left\{ (1 - q)\, e^{-D/R\theta} - \frac{\rho}{\rho_D} q^2 \right\} = \frac{k(1 + q)}{3} \theta_{,ii} \tag{7}$$

where

$$c_f^2 = \gamma(q) \frac{p}{\rho} \quad \text{and} \quad \gamma(q) = \frac{4 + q}{3}.$$

3. Velocity of propagation of sonic wave

We consider a moving singular surface $S(t)$ across which the temperature field and its first derivatives are continuous with possible discontinuities in the second derivatives while the other flow quantities are continuous with possible discontinuities in their first derivatives.

Let these discontinuities be denoted by

$$[v_{i,j}] = \lambda n_i n_j ; \quad \left[\frac{\partial v_i}{\partial t}\right] = -G\lambda n_i$$

$$[\rho_{,i}] = \zeta n_i ; \quad \left[\frac{\partial \rho}{\partial t}\right] = -G\zeta$$

$$[p_{,i}] = \xi n_i ; \quad \left[\frac{\partial p}{\partial t}\right] = -G\xi$$

where λ, ζ and ξ are scalars defined on $S(t)$.

Using first order compatibility conditions in (1)–(3) and (6), we get

$$(G - v_n)\,\zeta = \rho\lambda \tag{8}$$

$$\rho\,(G - v_n)\,\lambda = \xi \tag{9}$$

$$\tfrac{1}{3} K(1 + q)\,[\theta_{,ii}] = (c_f^2 - a_f^2)\,a_f \zeta \tag{10}$$

and

$$(G - v_n)\,[q_{,i}] = 0 \tag{11}$$

where $v_n = v_i n_i$;

$a_f^2 = \left(\frac{\partial p}{\partial \rho}\right)_{\theta,a}$; a_f is the frozen isothermal speed of sound.

Differentiating the state equation (4) with respect to x_i, taking jump across $S(t)$ and using (11), we get

$$\xi = a_f^2\,\zeta. \tag{12}$$

From equations (8), (9) and (12) we get

$$(G - v_n)^2 = a_f^2 \tag{13}$$

which indicates that the weak discontinuity in such a medium propagates with the frozen isothermal speed of sound.

If the medium infront of the surface $S(t)$ is uniform and at rest, it follows that the speed of propagation is constant and is given by $G = a_f$ and (8) and (9) can be written as

$$\xi = \rho G\lambda = G^2\,\zeta. \tag{14}$$

4. Fundamental differential equations for growth and decay of weak discontinuities in uniform medium

Differentiating the equations (1) and (2) with respect to x_i and using second order compatibility conditions we get

$$\frac{\delta\zeta}{\delta t} - G\zeta + \rho\bar{\lambda}_i n_i - 2\zeta\lambda - 2\rho\lambda\Omega = 0 \tag{15}$$

and

$$\rho \frac{\delta \lambda}{\delta t} + \bar{\xi} - \rho G \bar{\lambda}_i n_i = 0 \tag{16}$$

where

$$\bar{\lambda}_i = [v_{i,jk}]\, n_j n_k,$$

$$\bar{\xi} = [p_{,ij}]\, n_i n_j,$$

$$\zeta = [\rho_{,ij}]\, n_i n_j;$$

$\delta/\delta t$ denote differentiation along an orthogonal trajectory of the surface $S(t)$, and is given by $x_i = x_i(u_\alpha)$, $(\alpha = 1, 2)$; Ω is the mean curvature of $S(t)$ defined by $2\Omega = g^{\alpha\beta} b_{\alpha\beta}$, where $g^{\alpha\beta}$ and $b_{\alpha\beta}$ are the first and second fundamental forms of $S(t)$ respectively.

Now differentiation of equation (6) with respect to t and evaluation across $S(t)$ yields

$$(G - v_n)\, \bar{\mu} = P\zeta \tag{17}$$

where

$$P = C\theta^{-n} \left\{ \frac{2\rho}{\rho_D} q^2 - (1 - q)\, e^{-D/R\theta} \right\} \tag{18}$$

and

$$\bar{\mu} = [q_{,ii}].$$

Next, differentiating (4) twice with respect to x_i, x_j and taking jump, we get

$$\bar{\xi} = \frac{3\rho R}{K} (a_f C_f^2 - a_f^3) + a_f^2 \bar{\zeta} + \frac{Pp}{(1 + q)\, a_f} \zeta. \tag{19}$$

Eliminating $\bar{\xi}$, $\bar{\zeta}$ and $\bar{\lambda}_i$ from (15), (16) and (19) we obtained

$$\frac{\delta \lambda}{\delta t} - \lambda^2 - G\Omega\lambda + \left\{ \frac{3\rho R}{2k} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, a_f^2} \right\} \lambda = 0 \tag{20}$$

$$\frac{\delta \zeta}{\delta t} - \frac{G}{\rho} \zeta^2 - G\Omega\zeta + \left\{ \frac{3\rho R}{2k} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, a_f^2} \right\} \zeta = 0 \tag{21}$$

$$\frac{\delta \xi}{\delta t} - \frac{\xi^2}{G\rho} - G\Omega\xi + \left\{ \frac{3\rho R}{2k} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, a_f^2} \right\} \xi = 0 \tag{22}$$

or

$$\frac{d\lambda}{d\sigma} = \frac{\lambda^2}{G} + \left[\Omega - \left\{ \frac{3\rho R}{2KG} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, G^3} \right\} \right] \lambda \tag{23}$$

$$\frac{d\zeta}{d\sigma} = \frac{\zeta^2}{\rho} + \left[\Omega - \left\{ \frac{3\rho R}{2KG} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, G^3} \right\} \right] \zeta \tag{24}$$

$$\frac{d\xi}{d\sigma} = \frac{\xi^2}{\rho G^2} + \left[\Omega - \left\{ \frac{3\rho R}{2KG} (c_f^2 - a_f^2) + \frac{Pp}{2(1 + q)\, G^3} \right\} \right] \xi \tag{25}$$

while writing the equations (23)–(25) we have used the following relations:

$$\frac{\delta\lambda}{\delta t} = G\frac{d\lambda}{d\sigma}\,;\quad \frac{\delta\zeta}{\delta t} = G\frac{d\zeta}{d\sigma}\,;\quad \frac{\delta\xi}{\delta t} = G\frac{d\xi}{d\sigma} \tag{26}$$

where σ is the distance measured from $S(t_0)$ (the sonic surface at time t_0) along the normal trajectories to the family of surface $S(t)$ in the direction of propagation.

The equations (23), (24) and (25) constitute a set of basic differential equations for growth and decay of weak discontinuities associated with the wave $S(t)$ existing in thermally conducting and dissociative medium.

In the general case, Ω can be expressed in terms of the distance σ, given by the equation

$$\Omega = \frac{\Omega_0 - K_0\sigma}{1 - 2\Omega_0\sigma + K_0\sigma^2}$$

where K_0 and Ω_0 are the Gaussian and mean curvatures of the surface $S(t_0)$. Hence the integral of (25) by substitution of Ω from above equation, yields :

$$\frac{1}{\zeta} = \frac{1}{\zeta_0}(1 - 2\Omega_0\sigma + K_0\sigma^2)^{1/2}\, e^{Q\sigma}$$

$$+ \frac{1}{\rho}\, e^{Q\sigma}(1 - 2\Omega_0\sigma + K_0\sigma^2)^{1/2} \int_0^{\sigma} \frac{e^{-Q\sigma}}{(1 - 2\Omega_0\sigma + K_0\sigma^2)^{1/2}}\, d\sigma \tag{27}$$

where ζ_0 is the value of ζ at $\sigma = 0$ and

$$Q = \left\{\frac{3\rho R}{2KG}(c_f^2 - a_f^2) - \frac{Pp}{2(1+q)G^3}\right\}.$$

In order to predict physical situation more clearly we consider the particular cases of plane and spherical waves. For the case of plane wave, the equation (27) reduces to

$$\frac{1}{\zeta} = \frac{1}{\zeta_0}\left[e^{Q\sigma} + \frac{\zeta_0}{\rho Q}(e^{Q\sigma} - 1)\right]$$

or

$$\zeta = \frac{\zeta_0}{e^{Q\sigma} + \frac{1}{\rho Q}(e^{Q\sigma} - 1)\,\zeta_0}\,. \tag{28}$$

If ζ_0 is negative, the weak discontinuity grows continuously till it tends to infinity as

$$\sigma \to \frac{1}{Q}\ln\left\{\frac{|\zeta_0|}{|\zeta_0| - \rho Q}\right\}$$

and then sonic wave terminates into shock wave. The critical time t_c for the termination of compressive wave of order one into a shock is given by

$$t_c = t_0 + \frac{1}{QG}\ln\left\{\frac{|\zeta_0|}{|\zeta_0| - \rho Q}\right\}$$

It turns out that a weak discontinuity will grow and after a time t_o, it will terminate into a shock. The effect of dissociation and thermal conduction is to increase the critical time t_o. But if the weak discontinuity is an expansion wave of order 1, the discontinuity will continuously decay and will be damped out. The effect of dissociation and thermal conduction is to cause rapid damping. A similar remark is obtained if the singular surface $S(t)$ consists of a family of concentric spheres.

5. Non-uniform propagation of weak discontinuities

In order to discuss the non-uniform propagation of weak discontinuities, differentiate the equations (1) and (2) with respect to x_j, multiply them by n_j, sum over the index j and using the second order compatibility conditions we get the following equations :

$$U\frac{\delta\zeta}{\delta t} - (U^2\zeta - \rho U\bar{\lambda}_i n_i) + 2U\lambda\left[\left(\frac{\partial\rho}{\partial n}\right)_2 - \rho\Omega\right] - 2U\zeta\lambda + 2U\zeta\,(v_{i,j}\,n_i n_j)_2 + Ug^{\alpha\beta}\,v_i\zeta_{,\alpha}x_{i,\beta} = 0 \tag{29}$$

and

$$\rho\frac{\delta\lambda}{\delta t} + (\bar{\xi} - \rho U\bar{\lambda}_i n_i) - U\lambda\left(\frac{\partial\rho}{\partial n}\right)_2 + \left(\frac{\partial v_i}{\partial t} + v_k v_{i,k}\right)_2\zeta n_i + U\zeta\left(\frac{\partial v_k}{\partial n}\right)_2 n_k + U\zeta\,(v_{k,j}n_k n_j)_2 + \rho g^{\alpha\beta}\,v_i\lambda_{,\alpha}\,x_{i,\beta} = 0 \tag{30}$$

where symbols have their usual meaning and $U = (G - V_i u_i)$.

From equations (29) and (30) we write

$$\left(\frac{\delta\zeta}{\delta t} + v_i g^{\alpha\beta}\,\zeta_{,\alpha}x_{i,\beta}\right) + \rho\left(\frac{\delta\lambda}{\delta t} + v_i g^{\alpha\beta}\lambda_{,\alpha}x_{i,\beta}\right) + (\bar{\xi} - U^2\zeta) + 3U\zeta\,(v_{i,j}n_i n_j)_2 + U\zeta\left(\frac{\partial v_i}{\partial n}\right)_2 n_i - 2U\zeta\lambda + U\lambda\left(\frac{\partial\rho}{\delta n}\right)_2 - 2U\lambda\Omega + \left(\frac{\partial v_i}{\partial t} + v_k v_{i,k}\right)_2 n_i\zeta = 0. \tag{31}$$

Now differentiating the state equation (4), applying the second order compatibility conditions across $S(t)$ and using $[\theta_{,tt}] = \xi U/k$ and $\bar{\mu} = [q_{,ii}]$, we get

$$(\bar{\xi} - U^2\zeta) = \rho R\theta\bar{\mu} + 2R\theta\,(q_{,i})_2 n_i\zeta + 2R\,(1+q)\,(\theta_{,i})_2 n_i\zeta + \frac{U\rho R\,(1+q)}{K}\,\xi. \tag{32}$$

Next differentiating the rate equation (6) with respect to t and proceeding as above we write

$$(G - v_n)\,\bar{\mu} = \lambda\,(q_{,i})_2 n_i + P\zeta. \tag{33}$$

Substituting $\bar{\mu}$ from (33) in (32), we have

$$(\bar{\xi} - U^2\bar{\zeta}) = \frac{\rho R\theta}{U}\{\lambda\,(q_{,i})_2 n_i + P\zeta\} + 2R\theta\,(q_{,i})_2 n_i\zeta + 2R\,(1+q)\,(\theta_{,i})_2 n_i\zeta + \frac{U\rho R\,(1+q)}{R}\,\xi. \tag{34}$$

Eliminating $(\bar{\xi} - U^2\bar{\zeta})$ from (34) and (31) we arrive at the following equation

$$U\left(\frac{\delta\zeta}{\delta t} + v_i g^{\alpha\beta}\zeta_{,\alpha}x_{i,\beta}\right) + \rho\left(\frac{\delta\lambda}{\delta t} + v_i g^{\alpha\beta}\lambda_{,\alpha}x_{i,\beta}\right)$$
$$+ \frac{\rho R\theta}{U}(q_{,i})_2 n_i\lambda + \frac{\rho R\theta}{U}P\zeta + 2R\theta(q_{,i})_2 n_i\zeta + 2R(1+q)(\theta_{,i})_2 n_i\zeta$$
$$+ \frac{U\rho R(1+q)}{K}\xi + 3U\zeta(v_{i,j}n_i n_j)_2 + U\zeta\left(\frac{\partial v_i}{\partial n}\right)_2 n_i - 2U\zeta\lambda$$
$$+ U\lambda\left(\frac{\partial\rho}{\partial n}\right)_2 - 2\rho U\lambda\Omega + \left(\frac{\partial v_i}{\partial t} + v_k v_{i,k}\right)_2 n_i\zeta = 0. \tag{35}$$

This is a Riccatti differential equation for ζ along the orthogonal trajectories of $S(t)$ and is amenable to analysis. However the non-homogeneous terms arising from the surface derivatives cause some difficulty in integration and interpretation. But this difficulty is overcome by transforming it to a differential equation along the bicharacteristic curves in the characteristic manifold $\Sigma = \underset{t}{U} S(t)$ and finally we arrived at an ordinary differential equation (37).

Following Elcrat [6] we write

$$\frac{d\zeta}{dt} = \frac{\delta\zeta}{\delta t} + g^{\alpha\beta}v_i\zeta_{,\alpha}x_{i,\beta}$$
$$\frac{d\lambda}{dt} = \frac{\delta\lambda}{\delta t} + g^{\alpha\beta}v_i\lambda_{,\alpha}x_{i,\beta} \tag{36}$$
$$\frac{d\xi}{dt} = \frac{\delta\xi}{\delta t} + g^{\alpha\beta}v_i\xi_{,\alpha}x_{i,\beta}.$$

Using (8) and (36) in (35) we get

$$\frac{d\zeta}{dt} + \zeta\left\{\frac{1}{2}\frac{d}{dt}\ln\left(\frac{U}{\rho}\right) + \frac{1}{2U}E\right\} - \frac{U^2}{\rho}\zeta^2 = 0 \tag{37}$$

where

$$E = 3R\theta(q_{,i})_2 n_i + 2R(1+q)(\theta_{,i})_2 n_i + \frac{\rho R(1+q)}{K}U^3$$
$$+ \frac{\rho R\theta^{1-n}}{U}P + 3U(v_{i,j}n_i n_j)_2 + U\left(\frac{\partial v_i}{\partial n}\right)_2 n_i$$
$$+ \frac{U^2}{\rho}\left(\frac{\partial\rho}{\partial n}\right)_2 - 2U^2\Omega + \left(\frac{\partial v_i}{\partial t}\right) + v_k v_{i,k})_2 n_i.$$

This is the fundamental equation for the growth and decay of sonic discontinuities associated with the wave surface $S(t)$.

Integrating the equation (37), we get

$$\zeta = \frac{\zeta_0\left(\frac{U_0}{U}\right)^{1/2}\left(\frac{\rho}{\rho_0}\right)^{1/2}\exp\left(-\frac{1}{2U}\int_0^t E d\tau\right)}{1 - \frac{1}{2}\zeta_0 U_0^{\frac{1}{2}}\rho_0^{-\frac{1}{2}}\int_0^t\left(\frac{U}{\rho}\right)^{1/2}\exp\left(-\frac{1}{U}\right)\int_0^t E d\tau\, d\tau}$$

where $S(t) = S_0$, $\rho = \rho_0$, $U = U_0$, $\xi = \xi_0$, $\zeta = \zeta_0$ and $\lambda = \lambda_0$ at initial time $t = 0$.

If ζ_0 is positive, we have the criteria

$$\int_0^t U^{1/2} \rho^{-1/2} \exp\left(-\frac{1}{U}\int_0^t E d\tau\right) d\tau = \frac{2\rho_0^{\frac{1}{2}}}{\zeta_0 U_0^{\frac{1}{2}}}$$

for decay or "blow up" of the discontinuity at finite time T. If we associate $\zeta \to \infty$ with the formation of a shock, these remarks may be thought of as generalization of the corresponding statement given in § 4. Thus thermal conduction and dissociation allows the existence of a singular surface carrying a weak discontinuity into a non-uniform medium; this weak discontinuity grows into a shock, and the role of thermal conduction and dissociation is to cause damping in the formation of this shock.

Acknowledgement

The authors are grateful to Professor O P Bhutani for his constant encouragement and valuable suggestions and CSIR, Government of India, for financial assistance.

References

[1] Bazer J and Fleischman O 1959 *Phys. Fluid* **2** 366
[2] Bhutani O P and Rama Shankar 1970 *Indian Univ. Math. J.* **20** 239
[3] Clarke J F, Cleaver J W and Lilley G M 1964 *Symposium on Dissociating and ionizing gases in Engineering Inst. Mech. Engg. I.* (London : Birdcage Walk)
[4] Clarke J F 1969 *Small-disturbance theories, non-equilibrium flows Pt. I, Gasdynamics series* Ed. P P Wegener (New York and London: Marceland Dekker) **1** 1
[5] Duff G F D 1964 *Comm. Pure Appl. Math.* **1** 189
[6] Elcrat A E 1977 *Int. J. Engg. Sci.* **15** 29
[7] Kaul G N 1961 *J. Math. Mech.* **10** 393
[8] Landau L D and Lifshitz E M 1959 *Fluid mechanics* p. 259 (London : Pergamon Press)
[9] Lighthill M J 1960 *Phil. Trans. R. Soc.* **A 252** 397
[10] Ludwig D A 1961 *The singularities of the Riemann Function*, A.E.C. Report, NYO-9351, New York
[11] Luneberg R K 1964 *Mathematical theory of optics* (Berkeley : University of California Press)
[12] Nariboli G A 1963 *J. Math. Mech.* **12** 141
[13] Nariboli G A and Ranga Rao M P 1967 *J. Math. Phys. Sci.* **1** 302
[14] Nariboli G A, Singh S N and Ranga Rao M P 1968 *Proc. Indian Acad. Sci.* **A68** 302
[15] Rama Shankar and Chandran P 1977 *J. Math. Phys. Sci.* **11** 237
[16] Srinivasan S and Rishi Ram 1975 *Z. Angew. Math. Phys.* **26** 307
[17] Thomas T Y 1957 *J. Math. Mech.* **6** 311
[18] Thomas T Y 1957 *J. Math. Mech.* **6** 455
[19] Truesdell C A and Tourin R A 1960 *Handbuch der Physics* Ed. S Flugge Vol. III/I, (Berlin : Springer)
[20] Upadhyay K S 1970 *Tensor (N.S.)* **21** 296
[21] Varley E and Comberbatch E 1965 *J. Inst. Math. Its Appl.* **1** 101

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 61–66.

On the breakdown of acceleration waves in dissociating gas flows

RISHI RAM, BISHUN DEO PANDEY AND A S RAI
Department of Applied Mathematics, Institute of Technology, Banaras Hindu University, Varanasi 221 005

MS received 8 July 1978

Abstract. The present paper is devoted to the study of characteristic solution in the neighbourhood of the leading frozen characteristics in dissociating gas flows. It is found that at the cusp of the envelope of intersecting forward characteristics there occurs a breakdown of the wave after a finite critical time t_c. It is observed that there exists a critical value of the initial amplitude of the wave such that all compressive waves with an initial amplitude greater than the critical one will terminate into a shock wave due to non-linear steepening while an initial amplitude less than the critical one will result in a continuous decay. It is also concluded that the breakdown point moves forward along the leading characteristics due to dissociation effects.

Keywords. Non-linear; acceleration waves; shock waves; weak discontinuity; breakdown.

1. Introduction

Acceleration waves have been extensively studied during the last decade. The most interesting part of this study is that it is a subclass of non-linear waves which admit analytic solutions. Becker [1] and Bowen and Chen [2] studied various properties of acceleration waves in non-equilibrium flows. Varley [11] studied the growth and decay of acceleration waves in viscoelastic materials. Coleman and Gurtin [4] studied the local and global behaviour in materials with fading memory. Shuhubi and Jeffrey [10] studied the propagation of acceleration waves in layered hyperelastic half space. Ram and Srinivasan [8] studied the growth and decay of acceleration waves in radiating gases. Rarity [7] used Jeffrey and Taniuti's [5] technique to study the problem of breakdown of characteristic solutions in flows with vibrational relaxation. The main academic aspect of the present paper is to use Jeffrey and Taniuti's technique to investigate the dissociation effects and that of the wave geometry on the global behaviour of these waves during propagation. We assume a simple dissociating gas of Lighthill's model [6] under the temperature range 1000° K to 7000° K. In the temperature range where dissociation is important, the contribution of energy from electronic excitation and ionisation are both assumed negligible [3].

2. Dependence of the wave amplitude on time

Let us consider a symmetric transient gas flow with dissociation effects, which is induced by the motion of a piston advancing with finite acceleration into a gas at rest in the state of weak equilibrium. The constitutive equations by Ram and Gaur [9] governing the gas flow are

$$\frac{\partial \rho}{\partial t} + u\frac{\partial \rho}{\partial x} + \rho\frac{\partial u}{\partial x} + \delta\frac{\rho u}{x} = 0,$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{1}{\rho}\frac{\partial p}{\partial x} = 0,$$

$$\frac{\partial p}{\partial t} + u\frac{\partial p}{\partial x} + \gamma_e p\frac{\partial u}{\partial x} + F(p, \rho, a) = 0,$$

$$\frac{\partial a}{\partial t} + u\frac{\partial a}{\partial x} + \frac{3F(p, \rho, a)(1 + a)}{\rho D(1 + a)^2 - 3p} = 0,$$

where $F(p, \rho, a) = \dfrac{4\rho D^2 Kr}{3R^2 T_d^2}\{3p - \rho D(1 + a)^2\}\{\rho_d(1 - a)\exp(-T_d/T) - \rho a^2\}$.

and u, p, ρ, T represent the translation properties of the gas and have their usual meanings. a, D, k_r, T_d, ρ_d and R respectively denote the degree of dissociation, the dissociation energy per unit mass, the reaction rate constant, the characteristic temperature, the characteristic density and the gas constant. It has been observed that the variation of ρ_d over the temperature range for dissociation, i.e. from 10^3 °K to 7×10^3 °K is very slight. Hence for practical purposes the useful simplification of regarding ρ_d as a constant should lead to negligible errors. Here $\delta = 0, 1, 2$ for planar, cylindrical and spherical symmetry respectively and x is the radial distance from origin. $\gamma_e = (4 + a)/3$ is the effective exponent of heat for the gas mixture.

Now combining the basic equations, we have

$$U_t + AU_x + B = 0, \tag{1}$$

where U and B are the column vectors and A is a matrix of order 4, given by

$$U = \begin{bmatrix} p \\ u \\ \rho \\ a \end{bmatrix}, \quad A = \begin{bmatrix} u & \gamma_e p & 0 & 0 \\ \frac{1}{\rho} & u & 0 & 0 \\ 0 & \rho & u & 0 \\ 0 & 0 & 0 & u \end{bmatrix}, \quad B = \begin{bmatrix} F(p, \rho, a) \\ 0 \\ \delta\rho u/x \\ \dfrac{3F(p, \rho, a)(1 + a)}{\rho D(1 + a)^2 - 3p} \end{bmatrix}.$$

The system (1) is quasilinear and has four real characteristics along which acceleration waves propagate. The eigenvalues of A are

$$\lambda^1 = u + C, \ \lambda^2 = u - C, \ \lambda^3 = u = \lambda^4$$

where C is the effective local speed of sound. Corresponding eigenvectors are given by

$$L^1 = \begin{bmatrix} 1 \\ C\rho \\ C^2 \\ 0 \end{bmatrix}, \quad L^2 = \begin{bmatrix} 1 \\ -C\rho \\ C^2 \\ 0 \end{bmatrix}, \quad L^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad L^4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Introducing new curvilinear coordinates (ϕ, t') by the equations

$$\phi_t + \lambda^1 \phi_x = 0, \quad t = t' \tag{2}$$

which imply that $\phi(x, t) = 0$ is the leading forward characteristic front, if we level the family of characteristics by $\phi = t^*$, where t^* is the time when a wavefront is produced by the piston. Any flow property $f(x, t)$ is continuous across $\phi(x, t) = 0$ but $\partial f/\partial x$ and $\partial f/\partial t$ undergo finite jumps across it, such discontinuities are defined as "acceleration waves" advancing normal to themselves with the effective speed of sound relative to the gas flow. The Jacobian of the transformation is given by

$$J = x_\phi = 1/\phi_x. \tag{3}$$

Let us consider an open region R bounded by $\phi(x, t) = 0$ and $\xi(x, t) = 0$ such that no other characteristic emanating from the origin enters this region R. $f(x, t)$ remains smooth at least for a finite time throughout the region R except for boundaries (figure 1).

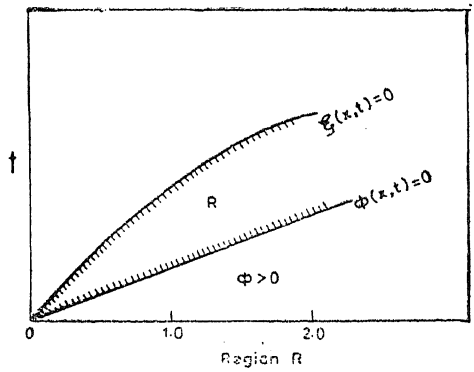

Figure 1. Region R.

Transforming (1) into new coordinate system and pre-multiplying by L^j, we have

$$L^j \left\{ x_\phi \frac{\partial}{\partial t'} + (\lambda - \lambda^1) \frac{\partial}{\partial \phi} \right\} U + x_\phi L^j B = 0 \tag{4}$$

which provides

$$L^1 (U_{t'} + B) = 0, \text{ for } \lambda = \lambda^1 \tag{5}$$

The boundary conditions are

$$[U_\phi]_{\phi=0^+}^{\phi=0^-} = \pi \neq 0, \quad [x_\phi]_{\phi=0^+}^{\phi=0^-} = X(t') \neq 0, \quad [U]_{\phi=0^+}^{\phi=0^-} = [U_{t'}]_{\phi=0^+}^{\phi=0^-} = 0$$

where $[x_\phi]_{\phi=0^+}^{\phi=0^-} = (x_\phi)_{\phi=0^-} - (x_\phi)_{\phi=0^+}$.

We assume that initially $\phi(x, t) = x$ at $t = 0$. Hence we observe that

$$X(t') + (x_\phi)_0 = (x_\phi)_{\phi=0^-}$$

is finite and non-zero, i.e. the transformation is non-singular if $X(t') + (x_\phi)_0$ is non-zero and finite. If U_0 corresponds to the constant state condition ahead of the wavefront, then $B(U_0) = 0$. It follows immediately from (4) that

$$L_0^\lambda \pi(t') = 0, \quad \lambda \neq \lambda^1. \tag{6}$$

Differentiating (5) with respect to ϕ at any point of the open region R and letting the point tend to the wavefront $\phi(x, t) = 0$, we get

$$L_0^1 \pi_{t'} + [\nabla_u (L^1 B)]_0 \pi = 0 \tag{7}$$

where ∇_u denotes the gradient operator with respect to the components of U.

The equation of the outgoing characteristics can be written as

$$x_{t'} = \lambda^1 = u + C. \tag{8}$$

Differentiating (8) with respect to ϕ at any point in R and letting the point tend to the wavefront, we get

$$X_{t'} = [\nabla_u (\lambda^1)]_0 \pi$$

which provides on integration that

$$X = \tilde{X} + \int_0^t [\nabla_u (\lambda^1)]_0 \pi \, dt', \tag{9}$$

where $\tilde{X} = \lim_{t' \to 0} X$.

We would expect the solution to breakdown after a finite critical time t_c determined by the intersection of the two outgoing characteristics emanating from the piston. At such a critical point we must have

$$X(t') + (x_\phi)_0 = 0,$$

so that t_c is given by

$$1 + \int_0^{tc} [\nabla_n (\lambda^1)]_0 \pi \mid \tilde{x}_\phi \, dt' = 0. \tag{10}$$

In consequence of (6) we have

$$\pi_3 = \pi_4 = 0, \quad \pi_1 = C_0 \rho_0 \pi_2. \tag{11}$$

Also we have

$$\pi_{2t'} + C_0 K(0) C_1 \pi_2 + \frac{C_2 C_0 K(0)}{(1 + C_0 K(0) t)} \pi_2 = 0, \tag{12}$$

where $C_1 = \left\{\frac{2\rho D^2 Kr}{3R^2 T_d^2 C_0 K(0)}\right\}_0 \{\rho_d \exp(-Td/T) - \rho a^2\}_0 \{3 - \rho D(1+a)^2\}_0$,

and $C_2 = \delta/2$,

are dimensionless constants of the constant state U_0 and $K(0)$ is the initial curvature of the wavefront. Thus the jump discontinuity

$$\pi = [U_\phi]_{\phi=0+}^{\phi=0-} \quad \text{across } \phi = 0$$

is dependent on time and can be expressed as

$$\pi = \tilde{\pi}_2 \exp(-C_1 C_0 K(0) t)(1 + C_0 K(0) t)^{-C_2} [C_0 \rho_0 \, 1 \, 0 \, 0]. \tag{13}$$

Now substituting from (13) into (10) we get

$$1 + \tilde{U}_s \frac{(\gamma_e + 2)}{2} \int_0^{t_c} \exp(-C_1 C_0 K(0) t)(1 + C_0 K(0) t)^{-C_2} dt \tag{14}$$

which determines the critical time t_c for an acceleration wave to breakdown and consequently a shock type discontinuity will appear. Now we define the amplitude of the wave as

$$a(t) = [u_x]_{\phi=0-}^{\phi=0+} = \pi_2 / x_\phi. \tag{15}$$

Using (9) and (13) in (15) we get

$$a(t) = \frac{(1 + C_0 K(0) t)^{-C_2} \exp(-C_1 C_0 K(0) t)}{1/a(0) + \frac{(\gamma e + 2)}{2} \int_0^{t_c} \exp(-C_1 C_0 K(0) t')(1 + C_0 K(0) t') dt} \tag{16}$$

which provides us the dependence of the wave amplitude $a(t)$ on time.

From the solution (16) we observe that the amplitude $a(t)$ for an expansion wave ($a(0) > 0$) will continuously decay with time and will vanish away ultimately. On the other hand for a compressive wave there exists a critical value a_c of the initial amplitude given by

$$a_c = \left\{\frac{(\gamma e + 2)}{2} \int_0^{\infty} \exp(-C_1 C_0 K(0) t')(1 + C_0 K(0) t') dt'\right\}^{-1}$$

such that if $|a(0)| > a_c$, there will occur a breakdown point on the leading wavefront and consequently a shock wave will be formed spontaneously, and if $|a(0)| < a_c$ there will be no breakdown.

3. Special case

Let us now confine our attention to the special case $C_2 = 0$ for which (16) reduces to the form

$$a(t) = \frac{a(0)}{(1 + a(0)/a_c)\exp(\mu t) - a(0)/a_c}, \tag{17}$$

where $\mu = \left(\frac{2\rho D^2 Kr}{3R^2 T_d^2}\right)_0 \left\{3\rho_d \exp(-T_d \mid T) - 3\rho a^2 + \rho^2 a^2 D(1+a)^2 - \rho_d D\rho (1+a)^2 \exp(-T_d \mid T)\right\}_0$

and $a_c = 2\mu \mid (\gamma e + 2)$,

is the critical value of the initial wave amplitude $a(0)$. From (17) it is clear that an acceleration wave is a compressive wave $(a(0) < 0)$ and $\mid a(0) \mid < a_c$, $a(t)$ increases exponentially and tends to zero as $t \to \infty$. If $\mid a(0) \mid$ coincides with a_c, the wave amplitude $a(t)$ becomes independent of time and is constant for all times. If $\mid a(0) \mid > a_c$, $a(t)$ increases exponentially and tends to infinity within a finite time t_e given by

$$t_e = \frac{1}{\mu} \log \frac{\mid a(0) \mid}{\mid a(0) \mid - a_c}$$

Thus when an initial wave amplitude exceeds the critical value a_c, there occurs a breakdown of the acceleration wave and consequently a shock wave will be formed. In case of an expansion wave, i.e. $a(0) > 0$ there occurs no breakdown and the wave will decay and will be damped out ultimately.

References

[1] Becker E 1970 *Aeronaut. J.* **74** 736
[2] Bowen R M and Chen P J 1972 *J. Math. Phys.* **13** 958
[3] Clarke J F 1960 *J. Fluid Mech.* **7** 577
[4] Coleman B D and Gurtin M E 1965 *Arch. Ration. Mech. Anal.* **19** 239
[5] Jeffrey A and Taniuti T 1964 *Non-linear wave propagation* (New York : Academic Press)
[6] Lighthill M J 1957 *J. Fluid Mech.* **2** 1
[7] Rarity B S H 1967 *J. Fluid Mech.* **27** 49
[8] Rishi Ram and Srinivasan S 1977 *ZAMM* **57** 191
[9] Rishi Ram and Gaur M 1976 *Acta Phys. Acad. Sci. Hung.* **40** 85
[10] Suhubi Erdogan S and Jeffrey A 1976 *Proc. R. Soc.* **75** 209
[11] Varley E 1965 *Arch. Ration. Mech. Anal.* **19** 215

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 1, January 1980, pp. 67–73.

Numerical solution of a quasilinear parabolic problem

P C JAIN and M K KADALBAJOO
Department of Mathematics, Indian Institute of Technology, Bombay 400 076

MS received 22 February 1979; revised 30 July 1979

Abstract. A combined approach of linearisation techniques and finite difference method is presented for obtaining the numerical solution of a quasilinear parabolic problem. The given problem is reduced to a sequence of linear problems by using the Picard or Newton methods. Each problem of this sequence is approximated by Crank-Nicolson difference scheme. The solutions of the resulting system of algebraic equations are obtained by using Block–Gaussian elimination method. Two numerical examples are solved by using both linearisation procedures to illustrate the method. For these examples, the Newton method is found to be more effective, especially when the given nonlinear problem has steep gradients.

Keywords. Parabolic equation; numerical method.

1. Introduction

Consider the quasilinear parabolic differential equation

$$u_t = u_{xx} + u_{yy} + \phi\,(x, y, t, u, u_x)\ ;\quad (x, y) \in R,\ 0 < t \leqslant T \tag{1}$$

subject to the initial condition

$$u\,(x, y, 0) = f(x, y),\ (x, y) \in R \tag{2}$$

and the boundary conditions

$$u\,(x, y, t) = g\,(x, y, t),\ (x, y) \in \partial R,\ 0 < t \leqslant T, \tag{3}$$

where R is a bounded plane region with a smooth boundary ∂R, and $f(x, y)$, $g\,(x, y, t)$ are given functions of their arguments defined in the respective domains. Existence and uniqueness theorems for the solution of this problem are known [5] under various assumptions on the functions ϕ, f and g. We assume that the conditions are satisfied so that the solution $u\,(x, y, t)$ exists and is unique with suitable regularity properties in the domain. Many authors [4, 6] have tackled this problem by replacing the partial derivatives by finite differences and solving the resulting system of linear or nonlinear equations by iterative methods. An excellent presentation of such methods is given in a survey paper by Douglas [3].

We have used combined approach of linearisation and finite difference method for solving this problem.

The problem is linearised by using Picard's method. This introduces a sequence of functions $\{u^{(n)}\}$ which satisfy the boundary conditions specified for u and linear partial differential equations

$$u_t^{(n+1)} = u_{xx}^{(n+1)} + u_{yy}^{(n+1)} + \phi\,[x, y, t, u^{(n)}, u_x^{(n)}], \quad (n = 0, 1, 2, \ldots), \tag{4}$$

$u^{(0)}$ being the initial guess.

If ϕ is a differentiable function, we can use Newton-linearisation which again introduces a sequence of functions $\{u^{(n)}\}$ that satisfy the boundary conditions specified for u and linear partial differential equations

$$u_t^{(n+1)} = u_{xx}^{(n+1)} + u_{yy}^{(n+1)} + \phi\,[x, y, t, u^{(n)}, u_x^{(n)}] + [u^{(n+1)} - u^{(n)}]\frac{\partial \phi}{\partial u} + [u_x^{(n+1)} - u_x^{(n)}]\frac{\partial \phi}{\partial u_x}, \quad (n = 0, 1, 2, \ldots), \tag{5}$$

($u^{(0)}$ being the initial guess), where the partial derivatives of ϕ are evaluated at the n-th step. When the sequence $\{u^{(n)}\}$ converges, the convergence is linear in the first case and quadratic in the second case.

The above sequence of linear problem equation (4) or (5) is then approximated by Crank–Nicolson difference scheme and the resulting algebraic problem is solved by Block–Gaussian elimination method. By using such a method, one can avoid the nonlinearity as well as the occurrence of two point boundary value problem. To be specific, we use the linearized version [equation (4)] of the given problem for the analysis of our method. It may be mentioned that the similar type of procedure can be applied to equation (5).

2. Crank–Nicolson difference scheme

Applying the Crank–Nicolson difference scheme to equation (4), we have

$$\left[\frac{u(x, y, t + \triangle) - u(x, y, t)}{\triangle}\right]^{(n+1)} = \tfrac{1}{2}[u_{xx}(x, y, t + \triangle) + u_{yy}(x, y, t + \triangle) + u_{xx}(x, y, t) + u_{yy}(x, y, t)]^{(n+1)} + \phi\,[x, y, t, \tfrac{1}{2}\{u(x, y, t + \triangle) + u(x, y, t)\}^{(n)}, \tfrac{1}{2}\{u_x(x, y, t + \triangle) + u_x(x, y, t)\}^{(n)}]. \tag{6}$$

For simplicity, we restrict our attention to the case where the given region R is a rectangle, $0 \leqslant x \leqslant a$, $0 \leqslant y \leqslant b$. We take $[a = Nh]$, $[b = Mh]$ (N, M positive integers) and space increments $h = \triangle x = \triangle y$. We use the notations

$$x_i = ih,\ y_j = jh,\ t_l = l\triangle,\ u_{i,j}(t_l) = u(ih, jh, l\triangle). \tag{7}$$

Replacing the partial derivatives in equation (6) by the standard finite difference approximations, we get the matrix-vector equation

$$u_i^{(n+1)}(t_{l+1}) = u_i^{(n)}(t_l) + r\,[u_{i+1}^{(n+1)}(t_{l+1}) - 2u_i^{(n+1)}(t_{l+1}) + u_{i-1}^{(n+1)}(t_{l+1}) - Qu_i^{(n+1)}(t_{l+1}) + u_{i+1}^{(n)}(t_l)$$

$$- 2u_i^{(n_l)}(t_l) + u_{i-1}^{(n_l)}(t_l) - Qu_i^{(n_l)}(t_l) + 2S_i]$$

$$+ \triangle\phi_i^{(n)}(t_{l+1}) \; ; \quad (i = 1,2,\ldots, N-1), \tag{8}$$

where the upper index n_l denotes the minimum number of iterations required to obtain an acceptable approximation to $u_{i,j}(t_l)$. The criterion for acceptance of an approximation will be discussed later when we study some test examples.

Here we have used the following notations :
The vectors

$$u_i(t_{l+1}) = \begin{bmatrix} u_{i,1}(t_{l+1}) \\ u_{i,2}(t_{l+1}) \\ \vdots \\ u_{i,M-1}(t_{l+1}) \end{bmatrix} ; \quad i = 1, 2, \ldots, N-1$$

$$S_i = [s_{i,j}] \; ; \text{ where } s_{i,j} = \begin{cases} u_{i,0} \; ; & j = 1 \\ u_{i,M} \; ; & j = M-1 \\ 0 \quad ; & \text{otherwise} \end{cases}$$

$$\phi_i(t_{l+1}) = \begin{bmatrix} \phi_{i,1}(t_{l+1}) \\ \phi_{i,2}(t_{l+1}) \\ \vdots \\ \phi_{i,M-1}(t_{l+1}) \end{bmatrix} ; \quad i = 1, 2, \ldots, N-1$$

with $r = \triangle/2h^2$.

The matrix

$$Q = (q_{ij}) \; ; \quad \text{where } q_{ij} = \begin{cases} 2 \; ; & i = j \\ -1 \; ; & |i-j| = 1 \\ 0 \; ; & \text{otherwise.} \end{cases}$$

Equation (8) is second order accurate in both space and time. It may be noted that the vector S_i remains the same for each time level due to the boundary conditions, and the vector $\phi_i(t_{l+1})$ is a known vector.

The matrix-vector equation (8) is solved by the Block–Gaussian elimination method [1].

3. Test examples

To test our method, we solve the following examples :
Example 1 :

$$u_t + uu_x = vu_{xx}; \; 0 < x < 1, \quad 0 < t \leqslant T, \tag{9}$$

subject to the initial and boundary conditions

$$u(x, 0) = \sin \pi x \; ; \; 0 \leqslant x \leqslant 1$$

$$u(0, t) = u(1, t) = 0; \; 0 < t \leqslant T. \tag{10}$$

Equation (9) is Burgers' equation well-known in the literature. We solve this problem by our method using both Picard as well as Newton linearisations. The linearised versions of equation (9) can be written as

$$u_t^{(n+1)} + u^{(n)} u_x^{(n)} = vu_{xx}^{(n+1)}, \qquad \text{(Picard)} \tag{11}$$

and

$$u_t^{(n+1)} + u^{(n)} u_x^{(n)} + [u^{(n+1)} - u^{(n)}] u_x^{(n)} + [u_x^{(n+1)} - u_x^{(n)}] u^{(n)} = \nu u_{xx}^{(n+1)}, \quad \text{(Newton)}, \tag{12}$$
$$(n = 0, 1, 2, \ldots)$$

subject to the initial and boundary conditions

$$u^{(n+1)}(x, 0) = \sin \pi x \; ; \; 0 \leqslant x \leqslant 1$$
$$u^{(n+1)}(0, t) = u^{(n+1)}(1, t) = 0 \; ; \; 0 < t \leqslant T. \tag{13}$$

We apply the Crank–Nicolson scheme and replace the partial derivatives by standard finite differences and express the solution of the resulting equations in the form

$$u_{i+1}^{(n+1)}(t_{l+1}) = a_i u_i^{(n+1)}(t_{l+1}) + b_i. \tag{14}$$

The solutions of this problem at certain grid points for different values of ν are given in tables 1 and 2. The criterion for the convergence was taken as

$$\underset{i}{\text{Max}} \mid u_i^{(n+1)}(t_{l+1}) - u_i^{(n)}(t_{l+1}) \mid \leqslant 10^{-6}.$$

Table 1. Numerical results for example 1 with $\nu = 0{\cdot}02$ and $\triangle = h = 0{\cdot}05$.

Mesh point (x,t)	Values of u	No. of iterations by Picard method	No. of iterations by Newton method
(0·05, 0·05)	0·134080		
(0·45, 0·05)	0·935674	10	4
(0·95, 0·05)	0·181590		
(0·15, 0·10)	0·308860		
(0·55, 0·10)	0·913854	20	3
(0·90, 0·10)	0·469373		
(0·15, 0·15)	0·233828		
(0·60, 0·15)	0·825115	32	4
(0·40, 0·20)	0·462650	43	4
(0·65, 0·20)	0·716397		
(0·25, 0·25)	0·225662	40	3
(0·35, 0·30)	0·247178	28	3
(0·50, 0·50)	0·158143	11	3
(0·65, 0·65)	0·112355	7	3
(0·65, 0·75)	0·075779	6	3
(0·45, 0·90)	0·037532	5	3
(0·70, 1·00)	0·024317	4	3
(0·60, 1·50)	0·001980	3	2
(0·50, 2·00)	0·000061	2	2

Table 2. Numerical results for example 1 with $\nu = 0{\cdot}005$ and $\triangle = h = 0{\cdot}05$.

Mesh point (x, t)	Values of u by Picard method	No. of iterations	Values of u by Newton method	No. of iterations
(0·05, 0·05)	0·134900		0·134900	
(0·50, 0·05)	0·975170	11	0·975170	4
(0·90, 0·05)	0·358865		0·358865	
(0·10, 0·10)	0·210120		0·210120	
(0·55, 0·10)	0·930457	57	0·930457	4
(0·85, 0·10)	0·673192		0·673192	
(0·15, 0·15)	*		0·237404	
(0·45, 0·15)	*		0·675632	
(0·85, 0·15)	*		0·866414	
(0·30, 0·30)	*		0·214687	
(0·70, 0·30)	*		0·493807	3
(0·50, 0·50)	*		0·160563	3
(0·75, 0·75)	*		0·115583	3
(0·75, 0·90)	*		0·079980	3
(0·60, 1·50)	*		0·021115	3
(0·70, 1·75)	*		0·012273	3
(0·50, 1·95)	*		0·008240	2

* Solution not converging

The number of iterations required for the convergence of solutions in both cases are also given in the tables. It may be noted that for $\nu = 1/200$ the scheme due to Picard linearization did not converge while the scheme using the Newton linearization yielded a solution in a few iterations.

Example 2 :

$$u_t - u_{xx} = (1 + u^2)\ (1 - 2u). \tag{15}$$

We consider it over two different triangles :

$$0 \leqslant t \leqslant 1 - x\ ;\ 0 \leqslant x \leqslant 1, \tag{16}$$

$$0 \leqslant t \leqslant 1{\cdot}5 - x\ ;\ 0 \leqslant x \leqslant 1{\cdot}5. \tag{17}$$

In each case, the boundary conditions are so chosen as to give the unique exact solution $u = \tan(x + t)$. This problem has earlier been studied by Bellman *et al* [2] by using iterative methods. We solve this problem by the proposed method.

The solutions at certain grid point for both the triangles are given in table 3. The convergence criterion was taken to be the same as in the previous example. In the case of smaller triangle, it is found that the execution time is the same for both Picard and Newton linearisations for the same accuracy. However, the situa-

Table 3. Numerical results for example 2 with $\triangle = h = 0{\cdot}05$.

Smaller triangle				Bigger triangle		
Mesh point (x, t)	Values of u	No. of iterations by Picard method	No. of iterations by Newton method	Mesh point (x, t)	Values of u by Newton method	No. of iterations
(0·05, 0·05)	0·100367			(0·10, 0·05)	0·151255	
(0·45, 0·05)	0·546670	6	3	(0·45, 0·05)	0·547423	
(0·90, 0·05)	1·398947			(0·95, 0·05)	1·578426	7
				(1·40, 0·05)	8·657083	
(0·10, 0·10)	0·221662			(0·10, 0·10)	0·223888	
(0·50, 0·10)	0·733277	7	4	(0·70, 0·10)	1·130229	7
(0·85, 0·10)	1·290839			(1·35, 0·10)	8·653630	
(0·30, 0·15)	0·550513			(0·80, 0·15)	1·579851	
(0·65, 0·15)	1·092695	8	4	(1·20, 0·15)	4·738682	7
(0·80, 0·15)	1·420196					
(0·25, 0·20)	0·477993	8	4	(0·50, 0·20)	1·001325	7
(0·50, 0·20)	0·917943			(1·00, 0·20)	2·807089	
(0·75, 0·20)	1·421981	8	4	(0·55, 0·30)	1·321651	7
(0·70, 0·25)	1·422444			(1·15, 0·30)	8·673449	
(0·40, 0·35)	0·997284	8	4	(0·50, 0·50)	1·756072	7
(0·55, 0·35)	1·300877					
(0·35, 0·50)	1·181929	7	4	(0·70, 0·75)	8·676374	7
(0·25, 0·70)	1·412587	6	3	(0·40, 1·00)	6·189992	7
(0·15, 0·80)	1·408046	5	3	(0·20, 1·25)	8·666633	7
(0·05, 0·90)	1·402138	4	3	(0·05, 1·35)	6·041356	6

tion is entirely different in the case of the bigger triangle, where only Newton linearisation is able to provide a solution. These observations are in agreement with those given by Bellman *et al* [2].

4. Discussion and conclusions

A study of the numerical results shows that the Newton method is more effective than the Picard method for obtaining the numerical solution of a quasilinear parabolic problem. The numerical results for the two examples considered have been given in tables 1–3.

Table 1 shows the results for the example 1 with $\nu = 1/50$ for both Picard and Newton linearisations. In this case, at the initial time levels the number of iterations taken by the Picard method are quite large as compared to the Newton method. However, this is not the case at the later time levels. This is due to the fact that

after a certain stage in time, the solutions approach towards steady states and the gradients become less steep as the time level goes on increasing. It is due to this reason that both Picard and Newton methods take approximately the same number of iterations in the vicinity of the steady state solution. For this example, the steady state solution can be considered to be attained at $t = 2{\cdot}00$. Table 2 gives the results for the same example, but with $\nu = 1/200$. In this case, Picard method provides solutions only upto the time level $t = 0{\cdot}10$. After this, the solution given by Picard method does not stabilise. On the contrary, the Newton method gives the solutions at all time levels by using a few iterations only. In this case, as ν decreases further, the convective term uu_x plays an important role.

The results for the second example are given in table 3. It is found that the execution time is the same for both the methods, although Picard method takes almost double the number of iterations compared to the Newton method; it may be due to the reason that the expressions to be computed by the Picard method are simpler than for the Newton case. The Newton method appears to have no advantage over the Picard method as there are no steep gradients in this case. However, the situation is entirely different for the same example considered in the bigger triangle. In this case, Picard method fails to yield a solution. This may be due to the reason that steep gradients occur in this triangle in which the boundary condition on $t = 1{\cdot}5 - x$ is $u = \tan 1{\cdot}5$ which is quite large. Based on these numerical results, one may conclude that the Newton method is more effective than the Picard method especially when the given nonlinear problem has steep gradients.

References

[1] Angel E and Bellman R 1972 *Dynamic programming and partial differential equations* (New York : Academic Press)

[2] Bellman R, Juncosa M and Kalaba R 1961 *Comm. ACM.* 4 187

[3] Douglas J Jr. 1961 *Advances in computers* 2 (ed) E L Alt (New York : Academic Press)

[4] Forsythe G E and Wasow W R 1960 *Finite difference methods for partial differential equations* (New York : John Wiley and Sons)

[5] Friedman A 1964 *Partial differential equations of parabolic type* (New Jersey : Prentice Hall)

[6] Richtmyer R D and Morton K W 1967 *Difference methods for initial value problems* (New York : Wiley-Interscience)

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 75–93.

A cryptographic system based on finite field transforms

E V KRISHNAMURTHY and VIJAYA RAMACHANDRAN
School of Automation, Indian Institute of Science, Bangalore 560 012, India

MS received 24 March 1979; revised 21 August 1979

Abstract. In this paper, we develop a cipher system based on finite field transforms. In this system, blocks of the input character-string are enciphered using congruence or modular transformations with respect to either primes or irreducible polynomials over a finite field. The polynomial system is shown to be clearly superior to the prime system for conventional cryptographic work.

Keywords. Ciphers; computer data protection; cryptography; finite field transforms; irreducible polynomials; military communication; polynomial congruences; prime congruences; public crypto-systems.

1. Introduction

A cryptographic system [7], [11] consists of a set of transformations T_j, each of which can act on an input message M to produce a corresponding enciphered message E_i, i.e.,

$$E_j = T_j(M).$$

Each transformation T_j is specified by an associated key k_j. The enciphered message is transmitted to the receiver over an interceptable medium. At the receiving end, the original message M is recovered by applying the inverse transformation T_j^{-1} on the received (enciphered) message E_j. It is clear that the existence of T_j^{-1} is a necessary condition for T_j to be a valid encoding transformation.

Example :

Substitution cipher—In this cipher, each input character is transformed into another character. The transformed character set may or may not be the same as the input character set. In the former case, the transformation is a simple permutation of the input character set. Let the (ordered) input characters be a, b, c, d. If the transformed character set is $\{0, 1, 2, 3\}$, any permutation of these 4 characters represents a key to a particular transformation. If 2, 1, 3, 0 is the specified key k_j, T_j is given by $a \to 2$, $b \to 1$, $c \to 3$, $d \to 0$, pna T_j^{-1} is given by $0 \to d$, $1 \to b$, $2 \to a$, and $3 \to c$. There are $4! = 24$ different transformations possible.

It is assumed that the enemy (i.e., the persons from whom the message M is protected) knows the set of transformations T_j being used. He also has available,

the *a priori* probabilities of the input characters. However, the enemy does not know which particular transformation is currently being used. This transformation T_j is characterised completely by the corresponding key k_j, which should be communicated to the receiving end through a protected channel.

If the enemy intercepts a sufficient number N of transmitted characters, he will be able to break the code, since he knows the *a priori* probabilities of the input characters. To illustrate this point, let $P(a) = 0{\cdot}8$, $P(b) = 0{\cdot}04$, $P(c) = 0{\cdot}15$, and $P(d) = 0{\cdot}01$ be the *a priori* probabilities in the example given above. Let the enemy intercept the 20 transmitted characters 3 2 2 2 1 2 0 2 2 2 2 3 3 2 2 2 2 2 2. As mentioned earlier, he knows that the cipher is one from the set of simple substitutions. By applying his knowledge of the *a priori* probabilities, he can immediately conclude that a is being coded as 2 and c as 3. At this stage he cannot assert anything about the encoding of b and d, but for a larger N he will be in a position to do so.

A good code (i.e., a set of encoding transformations) should satisfy two main criteria:

(i) N should become very large before the enemy is in a position to decipher the code using the *a priori* probabilities.

(ii) The number of elements in the key for the transformation should be relatively small.

Block coding is a very simple method of achieving (i). Here, the encoding transformation is applied to a block of k characters of the message at a time. If the number of different input characters is n, then block coding using k characters effectively increases the number of elements in the input character set to n^k. This is because each different sequence of k characters can be treated as an input character. In the example given above, if $k = 3$, the number of input characters for the block coding increases from 4 to $4^3 = 64$. The corresponding characters are 000, 001, 002, $\cdots$, 333. The probabilities for the occurrence of these characters can be calculated (with some difficulty); however, the value of N at which decoding by the enemy becomes possible is much larger than in the case of simple substitution. (It should be noted that the number of different keys in this case has increased to 64!). The price paid for this is the increase in the size of the key; while the key formerly contained 4 elements, it now contains 64 elements.

One attractive feature of the scheme presented in this paper is that it achieves block coding with a relatively small number of elements in the key. For instance, block coding with $k = 4$ and $n = 29$ can be achieved using just 5 elements in the key (as opposed to 29^4 elements for a block substitution code). Obviously, the number of different keys is much less than 29^4 !. However, the trade-off between key size and the number of transformations appears very attractive.

2. Finite structures from the number system

In this section, we shall outline some basic results from number theory and algebra. These results can be found from [9].

If p is a prime number, then the set of integers $0, 1, 2, \cdots, p-1$ together with addition and multiplication modulo p forms a field containing p elements. This is known as the *finite (Galois) field GF* (p). (It is easily verified that if p is not a

prime, then the resulting algebraic structure is not a field; it is a commutative ring).

Let

$$M = \prod_{i=1}^{k} p_i^{r_i}.$$

Then, given any positive integer a less than M, we can find its residues with respect to each $p_i^{r_i}$.

Let $a_i = a \bmod p_i^{r_i}$*, $0 \leqslant a_i < p_i^{r_i}$, $1 \leqslant i \leqslant k$.

To reconstruct a from the a_i's is not very straightforward. The formula for this reconstruction is given in the following theorem.

Theorem 1

Chinese remainder theorem. Let $p_1, p_2, \ldots, p_k$ be a set of k distinct primes and let $r_1, r_2, \ldots, r_k$ be positive integers. Let $a_1, a_2, \ldots, a_k$ be an arbitrary set of integers. Then the simultaneous congruences

$$a_i = a \bmod p_i^{r_i}, \quad i = 1, 2, \ldots, k$$

have a unique solution for a mod M, where

$$M = \prod_{i=1}^{k} p_i^{r_i}.$$

The solution is given by

$$a = \left(\sum_{i=1}^{k} d_i \left((a_i d_i^{-1}) \bmod p_i^{r_i}\right)\right) \bmod M,$$

where $d_i = M/p_i^{r_i}$, and $d_i^{-1} = (d_i \bmod p_i^{r_i})^{-1} \bmod p_i^{r_i}$.

(The inverse notation is defined in Definition 3 below).

Let a be a non-zero element of $GF(p)$. Then a^k is also a non-zero element of $GF(p)$ by the closure property. Since $GF(p)$ contains only p elements, there exist positive integers m and n such that

$$a^m = a^n \bmod p,$$

i.e., $a^k = 1 \bmod p$ for some integer $k > 0$.

Theorem 2. *Fermat's Theorem.* For every non-zero element a belonging to $GF(p)$,

$$a^{p-1} = 1 \bmod p.$$

Corollary 2.1. For every element a belonging to $GF(p)$, and any positive integer s,

$$a^{s(p-1)+1} = a \bmod p.$$

---

* For convenience, we use the equality sign for denoting congruence; this will not lead to any confusion as the statement modulo w.r.t. some integer will always follow the equality sign in such a case.

The extension of the result in Corollary 2.1 to the case when the modulus is the product of distinct primes is given in theorem 3. Before presenting theorem 3, however, we require the following two definitions.

Definition 1. *Euler's totient function* $\Phi(m)$. Let m be the positive integer. The number $\Phi(m)$ is defined to be the number of positive integers less than or equal to m that are relatively prime to m.

Let the prime decomposition of m be

$$p_1^{r_1} \cdot p_2^{r_2} \cdots p_k^{r_k}.$$

Then, $\Phi(m) = p_1^{(r_1-1)} \cdot p_2^{(r_2-1)} \cdots p_k^{(r_k-1)} (p_1 - 1)(p_2 - 1) \cdots (p_k - 1)$.

In the case when m is the product of distinct primes (i.e., $m = p_1 p_2 \cdots p_k$), $\Phi(m) = (p_1 - 1)(p_2 - 1) \cdots (p_k - 1)$.

Definition 2. Given two integers a and b, the notation (a, b) defines the greatest common divisor (*gcd*) of a and b. We define $(O, b) = b$.

Theorem 3. Let m be the product of distinct primes $p_1, p_2, \cdots, p_k$. Then, for every integer a, and any positive integer s,

$$a^{s\Phi(m)+1} = a \bmod m.$$

Proof. Let $a_i = a \bmod p_i$, $i = 1, 2, \cdots, k$.

Case 1. $(a,m) = 1$

Then $(a_i, p_i) = 1, i = 1, 2, \cdots, k$.

Hence $a_i^{p_i-1} = 1 \bmod p_i$ by Theorem 2

i.e., $a^{r(p_i-1)} = 1 \bmod p$

Hence $a^{p_i-1} = 1 \bmod p_i$.

Thus $a^{s\Phi(m)} = 1 \bmod m$, since $\Phi(m) = \prod_{i=1}^{k} (p_i - 1)$

or $a^{s\Phi(m)+1} = a \bmod m$.

Case 2. $(a, m) > 1$

Without loss of generality, we can assume

$$a = c \cdot p_1 \cdot p_2 \cdots p_j, \text{ and } (a, p_i) = 1, \quad j < i \leqslant k.$$

Then $a_i = 0$ for $1 \leqslant i \leqslant j$ and $a_i^{p_i-1} = 1 \bmod p_i$, $j < i \leqslant k$. (1)

Thus $a^{r(p_i-1)+1} = a \bmod p_i$, $j < i \leqslant k$ (2)

Also $a^r = 0 \bmod p_i$ for any integer r, $1 \leqslant i \leqslant j$. (3)

Now consider $a^{s\Phi(m)+1}$.

By (1), (2), and (3) we have

$$a^{s\Phi(m)+1} = a_i \bmod p_i, \quad 1 \leqslant i \leqslant k,$$

or $a^{s\Phi(m)+1} = a \mod m$.

We now present a few standard theorems from the theory of congruences.

Theorem 4. If $(a, m) = 1$, then the congruence $ax = b \mod m$ has a unique solution $x = x_1$ with $0 \leqslant x_1 < m$.

Definition 3. Let $(a, m) = 1$. Then, b is the inverse of a mod m (denoted a^{-1}) if $ab = 1 \mod m$.

Theorem 5. Let p be a prime and let $(n, p-1) = 1$. Then the congruence

$$x^n = a \mod p$$

has exactly one solution given by

$$x = a^{m_0} \mod p,$$

where $m_0 = n^{-1} \mod p - 1$.

Theorem 6. Let m be the product of distinct primes $p_1, p_2, \cdots, p_k$ and let $(n, \Phi(m)) = 1$. Then the congruence

$$x^n = a \mod m$$

has exactly one solution

$$x = a^{m_0} \mod m$$

where $m_0 = n^{-1} \mod \Phi(m)$.

Proof

$$a^{m_0} = x^{nm_0} \mod m.$$

But $nm_0 = 1 \mod \Phi(m)$

or $nm_0 = s \cdot \Phi(m) + 1$ for some positive integer s.

Thus, $a^{m_0} = x^{s\Phi(m)+1} \mod m$

$= x \mod m$ by theorem 3.

3. Polynomial systems

In this section, we consider finite fields and rings generated by polynomials over $GF(p)$ [3], [4].

Definition 4. A polynomial is said to be *defined* over $GF(p)$ if its coefficients lie in $GF(p)$.

Definition 5. A polynomial $\phi(x)$ is an *irreducible polynomial* if it has no divisors other than scalars and scalar multiples of itself.

Let $\phi(x)$ be an irreducible polynomial of degree d over $GF(p)$. Then the set of all polynomials over $GF(p)$ with degree less than d, together with addition and multiplication modulo $\phi(x)$, forms a field. This field is known as the *finite (Galois)*

field $GF(p^d)$ and it contains p^d elements. The additive identity is 0 and the multiplicative identity is 1. If $\phi(x)$ is not irreducible, then the above structure becomes a finite commutative ring with the same identity elements.

Definition 6. Let $a(x)$ belong to $GF(p^d)$ and let $\phi(x)$ be an irreducible polynomial over $GF(p)$ of degree d. Then, the least positive integer k for which the equation

$$a^k(x) = 1 \bmod \phi(x)$$

is satisfied, is called the *order* of $a(x)$.

Definition 7. If $k = p^d - 1$, then $a(x)$ is called a *primitive element* of the above field.

The powers of a primitive element generate all the non-zero elements of the finite field.

Given two irreducible polynomials $\phi(x)$ and $\psi(x)$ of degree d over $GF(p)$, we can generate two different fields contaning p^d elements by taking the field operations modulo $\phi(x)$ and $\psi(x)$ respectively.

Theorem 7. The number of irreducible polynomials of degree d over $GF(p)$ is given by

$$I_{d,p} = \frac{1}{d} \sum_{\substack{k, \\ k \mid d}} \mu(k) \cdot p^{d/k},$$

where $\mu(k) = \begin{cases} 1 \text{ if } k = 1 \\ (-1)^r \text{ if } k \text{ is the product of } r \text{ distinct primes} \\ 0 \text{ if } k \text{ contains any repeated prime factors.} \end{cases}$

Thus, we can have $I_{d,p}$ different representations of $GF(p^d)$ by choosing different irreducible polynomials.

We now generalise some of the results of the previous section to the polynomial case.

Theorem 8. *Chinese remainder theorem for polynomials.* Let $\phi_1(x), \phi_2(x), \cdots, \phi_k(x)$, be irreducible polynomials over $GF(p)$ and let $r_1, r_2, \cdots, r_k$ be positive integers. Let $a_1(x), a_2(x), \cdots, a_k(x)$ be arbitrary polynomials over $GF(p)$. Then, the simultaneous congruences

$$a_i(x) = a(x) \bmod \phi_i^{r_i}(x), \quad i = 1, 2, \cdots, k$$

have a unique solution for $a(x)$ modulo $\phi(x)$ where

$$\phi(x) = \prod_{i=1}^{k} \phi_i^{r_i}(x).$$

The solution is given by

$$a(x) = \left(\sum_{i=1}^{k} d_i(x) \left[(a_i(x)\, d_i^{-1}(x)) \bmod \phi_i^{r_i}(x) \right] \right) \bmod \phi(x),$$

where $d_i(x) = \phi(x)/\phi_i^{r_i}(x)$

and $\quad d_i^{-1}(x) = (d_i(x) \bmod \phi_i^{r_i}(x))^{-1} \bmod \phi_i^{r_i}(x).$

Theorem 9. Let $\phi(x)$ be an irreducible polynomial over $GF(p)$ of degree d. Then, for every non-zero $a(x)$ in $GF(p^d)$,

$$a^{p^d-1}(x) = 1 \bmod \phi(x).$$

Corollary 9.1. For every $a(x)$ in $GF(p^d)$,

$$a^{s\cdot(p^d-1)+1}(x) = a(x) \bmod \phi(x),$$

where s is any positive integer.

Before we extend the above result to the case of composite modulus, we introduce the concept of ' generalised totient function '.

Definition 8. *Generalised totient function*

Let $\phi(x)$ be a polynomial over $GF(p)$ which is the product of distinct irreducible polynomials $\phi_1(x), \phi_2(x), \cdots, \phi_k(x)$. Let $\delta_1, \delta_2, \cdots, \delta_j$ be distinct positive integers representing the degrees of the ϕ_i's. Clearly $j \leqslant k$. We define the generalised totient function of $\phi(x)$ as

$$Q(\phi) = \prod_{i=1}^{j} (p^{\delta_i} - 1).$$

Note. $Q(\phi)$ is not the ' natural ' extension of the number-theoretic totient function, in the sense that it does not give the number of polynomials over $GF(p)$ of degree less than $\phi(x)$ and relatively prime to it. However, this is the generalisation which is useful for the extension of Corollary 9.1.

Theorem 10. Let $\phi(x)$ be a polynomial of degree d over $GF(p)$, which is the product of distinct irreducible. Polynomials $\phi_1(x), \phi_2(x), \cdots, \phi_k(x)$. Let $\delta_1, \delta_2, \cdots, \delta_j$ be distinct positive integers, representing the degrees of the ϕ_i's. Then, for every polynomial $a(x)$ over $GF(p)$ of degree less than d,

$$a^{s\cdot Q(\phi)+1}(x) = a(x) \bmod \phi(x),$$

where s is any positive integer.

Proof. Similar to the proof of theorem 3.

Theorem 11. Let $\phi(x)$ be an irreducible polynomial of degree d over $GF(p)$ and let $(n, p^d - 1) = 1$. Then the congruence

$$w^n(x) = a(x) \bmod \phi(x)$$

has exactly one solution, given by

$$w(x) = a^{m_0}(x) \bmod \phi(x),$$

where $\quad m_0 = n^{-1} \bmod (p^d - 1).$

Proof. Let $\alpha(x)$ be a primitive element of the finite field defined by $\phi(x)$ and let $\alpha^j(x) = a(x) \bmod \phi(x)$ for some $j, 0 < j \leqslant p^d - 1$.

Now, let $w(x) = a^k(x) \bmod \phi(x)$ satisfy the equation

$$w^n(x) = a(x) \bmod \phi(x),$$

i.e., $$a^{kn}(x) = a^j(x) \bmod \phi(x).$$

Hence, $nk = j \bmod (p^d - 1)$ by Definition 10.

By theorem 5, this equation has exactly one solution for k if $(n, p^d - 1) = 1$, which is true by assumption.
Thus $k = jn^{-1} \bmod (p^d - 1)$,

or $$w(x) = a^{jn^{-1}}(x) \bmod \phi(x) = a^{n^{-1}}(x) \bmod \phi(x)$$

$$= a^{m_0}(x) \bmod \phi(x).$$

Theorem 12. Let $\phi(x)$ be the product of distinct irreducible polynomials and let $(n, Q(\phi)) = 1$. Then the congruence

$w^n(x) = a(x) \bmod \phi(x)$ has exactly one solution, given by

$w(x) = a^{m_0}(x) \bmod \phi(x)$, where $m_0 = n^{-1} \bmod Q(\phi)$.

Proof. Similar to the proof of theorem 6.

The proofs for theorems stated in this section without proof can be found in [3] and [4].

4. Prime congruence codes

In this section we describe the congruence system for primes, using the results of § 2. The input character set consisting of K symbols, is first converted into a set of k-bit codes (substitution codes), where k is the smallest number satisfying the inequality $2^k > K$. The input string is processed m characters at a time. The string $c_1, c_2, \ldots, c_m$ is interpreted as the number

$$M = c_1 + c_2 \cdot 2^k + \ldots + c_m \cdot 2^{k \cdot (m-1)}.$$

Hardware-wise, this represents the simple concatenation of the binary representation of the m characters.

A set of r distinct primes $p_1, p_2, \ldots, p_r$ is chosen such that $P = p_1 p_2 \cdots p_r > 2^{mk}$. Also, a positive integer n is chosen such that

$$(n, p_i - 1) = 1, \; i = 1, 2, \ldots, r, \; 1 < n < P',$$

where $$P' = \prod_{i=1}^{r} (p_i - 1).$$

The set of r primes p_i together with n constitute the key to the cipher. The encoding is done by taking the residue x_i of M with respect to each prime p_i, determining the e_i's where

$$e_i = x_i^{n_i}, \; i = 1, 2, \ldots, r,$$

and $$n_i = n \bmod (p_i - 1),$$

and combining the e_i's by the Chinese Remainder Theorem to obtain E. By Corollary 2.1, $x_i^n = x_i^{n_i}$. Hence,

$$E = M^n \bmod P,$$

and this is the enciphered message which is transmitted.

At the receiving end, the code is deciphered in an analogous manner as follows. The residue e_i of the message E with respect to each prime p_i is determined, viz.,

$$e_i = E \bmod p_i, \quad i = 1, 2, \ldots, r.$$

By theorem 6, the unique solution x_i to the congruence

$$x_i^n = e_i \bmod p_i, \quad i = 1, 2, \ldots, r$$

can be determined. The individual x_i's are then recombined using an efficient implementation of theorem 1 (Chinese Remainder Theorem) to obtain M.

5. Polynomial congruence codes

We now consider the polynomial congruence coding scheme. A prime p is chosen to satisfy the inequality $p > K$, where K is the number of symbols in the input character set. The number p would normally be the smallest prime larger than K. A substitution code is prepared, which maps the input character set into a set of distinct elements in $GF(p)$.

The input string is processed m characters at a time. A set of r irreducible polynomials over $GF(p), \phi_1(x), \phi_2(x), \ldots, \phi_r(x)$ is chosen, with degrees $d_1, d_2, \ldots, d_r$, such that

$$\sum_{i=1}^{r} d_i = m.$$

Let $\phi(x) = \prod_{i=1}^{r} \phi_i(x)$. An integer n is chosen,

satisfying the property

$$(n, p^{d_i} - 1) = 1, \quad i = 1, 2, \ldots, r, \quad 1 < n < Q(\phi).$$

Let $M(x)$ be the polynomial obtained by treating the m characters of the input string as coefficients of successive powers of x. The encoding is achieved by taking the residue $w_i(x)$ of $M(x)$ with respect to each $\phi_i(x)$, determining $e_i(x)$, where

$$e_i(x) = w_i^{n_i}(x), n_i = n \bmod (p^{d_i} - 1), i = 1, 2, \ldots, r,$$

and combining the $e_i(x)$ by the Chinese Remainder Theorem to obtain $E(x)$. It is clear that

$$E(x) = M^n(x) \bmod \phi(x).$$

The key to this cipher consists of n and the set of polynomials, $\phi_i(x)$, $i = 1, 2, \ldots, r$.

The receiver decodes the cipher in an analogous manner as follows. The residues $e_i(x)$ of the transmitted message $E(x)$ with respect to each $\phi_i(x)$ are

obtained. By theorems 11 and 12, the unique solution $w_i(x)$ to the congruence

$$w_i^n(x) = e_i(x) \bmod \phi_i(x), \quad i = 1, 2, \ldots, r$$

can be determined. The individual $w_i(x)$ thus obtained are combined using an efficient implementation of theorem 8 to obtain $M(x)$.

Encoding and decoding algorithms

Algorithm 1: *Encoding*

Input

(a) The block length m and the value of the exponent n.

(b) The r distinct irreducible polynomials over $GF(p)$, $\phi_1(x), \phi_2(x), \ldots, \phi_r(x)$ of degrees $d_1, d_2, \ldots, d_r$ respectively. The block length

$$m = \sum_{i=1}^{r} d_i.$$

(c) A substitution code table (τ) for the input character set. Each coded value lies in $GF(p)$.

(d) The input character string.

Output. The coded message string.

Algorithm

begin$_0$
  for $i \leftarrow 1$ *until* r *do*
    begin$_1$
2. $n_i \leftarrow n \bmod (p^{d_i} - 1)$
    end$_1$
end$_0$
begin$_2$
3. *while* input string is present *do*
    begin$_3$
4. read in the next m characters, $s_1, s_2, \ldots, s_m$
5. form the corresponding substitution code $c_1, c_2, \ldots, c_m$ by table look-up (using table τ)
6. $M(x) \leftarrow c_1 + c_2 x + \ldots + c_m x^{m-1}$
7. *for* $i \leftarrow 1$ *until* r *do*
      begin$_4$
8. $w_i(x) \leftarrow M(x) \bmod \varphi_i(x)$
9. $e_i(x) \leftarrow w_i^{n_i}(x) \bmod \varphi_i(x)$
      end$_4$

10. combine $e_1(x), e_2(x), \ldots, e_r(x)$ by Theorem 8 to obtain $E(x)$
11. write the coefficients of $E(x)$, starting with the constant term

end$_3$

end$_2$

Algorithm 2. Decoding

Input

(a) The block length m and the value of the exponent n.

(b) The coefficients of the r distinct irreducible polynomials $\varphi_1(x), \varphi_2(x), \ldots, \varphi_r(x)$ over $GF(p)$ of degrees $d_1, d_2, \ldots, d_r$ respectively.

(c) A table (τ') for back conversion from the $GF(p)$ codes to the message character set.

(d) The coded input.

Output. The decoded message string.

Algorithm

begin$_0$

1. *for* $i \leftarrow 1$ *until* r *do*

begin$_1$

2. $m_i \leftarrow n^{-1} \bmod (p^{a_i} - 1)$

end$_1$

end$_0$

begin$_2$

3. *while* input string is present *do*

begin$_3$

4. read in the next m characters (i.e., the coefficients of $E(x)$)
5. *for* $i \leftarrow 1$ *until* r *do*

begin$_4$

6. $e_i(x) \leftarrow E(x) \bmod \varphi_i(x)$
7. $w_i(x) \leftarrow e_i^{m_i}(x) \bmod \varphi_i(x)$

end$_4$

8. combine $w_1(x), w_2(x), \ldots, w_r(x)$ using Theorem 8 to obtain $M(x)$
9. decode the coefficients of $M(x)$ by table look-up (using table τ')
10. write the decoded version of the coefficients of $M(x)$, starting with the constant term

end$_3$

end$_2$

Example

We use the block size of $m = 4$ and proceed to illustrate Algorithms 1 and 2 using the message

TAT TVAM ASI

We use $GF(29)$ since this will conveniently accommodate the English alphabet, along with full-stop and blank. The substitution code used for this example is given in table 1.

We obtain the block-length of 4 by using two irreducible polynomials of degree 2 over $GF(29)$.

$$\varphi_1(x) = 8 + 10x + x^2,$$

$$\varphi_2(x) = 2 + 7x + x^2.$$

Hence $\varphi(x) = \varphi_1(x) \cdot \varphi_2(x) = 16 + 4x + 17x^2 + x^3$.

The exponent n can take any value between 1 and $Q(\varphi)$ that is relatively prime to $Q(\varphi)$. We shall use $n = 517$. The number of different values for n is $\Phi(Q(\varphi)) - 1 = \Phi(29^2 - 1) - 1 = 191$. Thus, the number of different keys in this system is

$$\binom{I_{29,2}}{2} \cdot 191.$$

$$I_{29,2} = \tfrac{1}{2} \sum_{k,\, k|d} \mu(k) \cdot 29^{d/k}$$

$$= \tfrac{1}{2} \cdot (\mu(1) \cdot 29^2 + \mu(2) \cdot 29)$$

$$= \tfrac{1}{2} \cdot (1 \cdot 841 - 29) = 406,$$

Table 1. Substitution code for example.

Character	Code	Character	Code
A	14	O	15
B	28	P	2
C	27	Q	4
D	25	R	8
E	21	S	0
F	13	T	16
G	26	U	3
H	23	V	6
I	17	W	12
J	5	X	24
K	10	Y	19
L	20	Z	9
M	11	·	18
N	22	b	7

i.e., the number of different keys

$$= 191 \frac{406 \times 405}{2} = 15703065.$$

The encoding procedure is shown in table 2.
The transmitted message string is

4 9 12 4 13 22 10 2 4 23 13 13

At the receiving end, this string is again processed 4 symbols at a time. The decoding steps are shown in table 3. At the end of the procedure, the decoded message is obtained as

TAT TVAM ASI

which is the message originally transmitted.

The key for this scheme consists of the following 5 elements: $n = 517$, constant term of $\varphi_1(x) = 8$, coefficient of x in $\varphi_1(x) = 10$, constant term of $\varphi_2(x) = 2$, and coefficient of x in $\varphi_2(x) = 7$.

6. System performance

In this section, we compare the performance of the procedures presented in § 4 and § 5. The following points should be noted:

(a) *Number of keys* : In the prime coding scheme, the number of keys corresponding to a given set of r primes is given by $\Phi(P') - 1$, where

$$P' = \prod_{i=1}^{r} (p_i - 1).$$

If k such sets of primes are chosen for the system, the total number of different keys is given by

$$\sum_{j=1}^{k} (\Phi(P'_j) - 1),$$

Table 2. An example of encoding using algorithm 1. Block length $m = 4$; Prime field: $GF(29)$; Number of irreducible polynomials $r = 2$; $\phi'(x) = 8 + 10x + x^2$; $\phi_2(x) = 2 + 7x + x^2$; Exponent $n = 517$; Input message: TAT TVAM ASI.

Block Number	Input Block	$M(x)$	$w_i(x) = M(x) \bmod \phi_i(x)$	$e_i(x) = w_i^n(x) \bmod \phi_i(x)$	$E(x)$
1	TAT␣	$16 + 14x + 16x^2 + 7x^3$	$w_1(x) = 13 + 5x$ $w_2(x) = 24 + 28x$	$e_1(x) = 25 + 25x$ $e_2(x) = 7 + 26x$	$4 + 9x + 12x^2 + 4x^3$
2	TVAM	$16 + 6x + 14x^2 + 11x^3$	$w_1(x) = 1 + 8x$ $w_2(x) = 26 + 19x$	$e_1(x) = 6 + 19x$ $e_2(x) = 21 + 17x$	$13 + 22x + 10x^2 + 2x^3$
3	␣ASI	$7 + 14x + 17x^3$	$w_1(x) = 4 + 12x$ $w_2(x) = 13 + x$	$e_1(x) = 12 + 16x$ $e_2(x) = 15 + 21x$	$4 + 23x + 13x^2 + 13x^3$

Table 3. An example of decoding using algorithm 2.

Block length $m = 4$; Prime field: $GF(29)$; Number of irreducible polynomials $r = 2$; $\phi_1(x) = 8 + 10x + x^2$; $\phi_2(x) = 2 + 7x + x^2$; Exponent $n = 517$; $m_0 = n^{-1} \bmod 840 = 13$; Input string. 4 9 12 4 13 22 10 2 4 23 13 13.

Block Number	$E(x)$	$e_i(x) = E(x) \bmod \phi_i(x)$	$w_i(x) = e_i^{m_0}(x) \bmod \phi_i(x)$	$M(x)$	Message string
1	$4 + 9x + 12x^2 + 4x^3$	$e_1(x) = 25 + 25x$ $e_2(x) = 7 + 26x$	$w_1(x) = 13 + 5x$ $w_2(x) = 24 + 28x$	$16 + 14x + 16x^2 + 7x^3$	TATb
2	$13 + 22x + 10x^2 + 2x^3$	$e_1(x) = 6 + 19x$ $e_2(x) = 21 + 17x$	$w_1(x) = 1 + 8x$ $w_2(x) = 26 + 19x$	$16x + 6x + 14x^2 + 11x^3$	TVAM
3	$4 + 23x + 13x^2 + 13x^3$	$e_1(x) = 12 + 16x$ $e_2(x) = 15 + 21x$	$w_1(x) = 4 + 12x$ $w_2(x) = 13 + x$	$7 + 14x + 17x^3$	bASI

where the extension of the earlier notation to cover different sets of primes is obvious.

In the polynomial case, let us assume, as before, that we have r irreducible polynomials over $GF(p)$ of degrees $d_1, d_2, \ldots, d_r$ respectively, whose product is $\varphi(x)$. Let $\delta_1, \delta_2, \ldots, \delta_k$ be distinct positive integers representing the degrees of the φ_i's. Further, let there be r_1 irreducible polynomials of degree δ_1, r_2 irreducible polynomials of degree δ_2, $\ldots$, r_k irreducible polynomials of degrees δ_k. Clearly, $\sum_{i=1}^{k} r_i = r$. Then, the number of different keys in the system is immediately determined as

$$(\Phi(Q(\phi)) - 1) \cdot \prod_{j=1}^{k} \binom{I_{\delta_j, p}}{r_j}.$$

We arrive at this number as follows: The number of irreducible polynomials of degree δ_j over $GF(p)$ is $I_{\delta_j, p}$ (by theorem 7). The number of different ways in which we can choose r_j polynomials from this set is

$$\binom{I_{\delta_j, p}}{r_i}.$$

By theorem 12, the number of permitted values of n is $\Phi(Q(\varphi)) - 1$. Hence the above formula gives the number of different keys when irreducible polynomials are used.

(b) *Exponentiation* : A fast algorithm for the computation of $M^n \bmod P$ is available in Knuth (1969, p. 399) [8]. The algorithm is given below.

Algorithm 3. *Exponentiation*

Input. A modulus P, a positive integer $M < P$, and a positive exponent n.

Output. $M^n \bmod P$.

Algorithm

begin$_0$

1. *initialise* $N \leftarrow n$; $Y \leftarrow 1$; $Z \leftarrow M$
2. *while* $N \neq 1$ *do*
   begin$_1$
3. *if* N is odd *then*
   begin$_2$
4. $Y \leftarrow Z \cdot Y \bmod P$
5. $Z \leftarrow Z \cdot Z \bmod P$
   end$_2$
6. *else* $Z \leftarrow Z \cdot Z \bmod P$
7. $N \leftarrow [N/2]$

end$_1$

8. $Y \leftarrow Y \cdot Z \bmod P$
9. *write* Y

end$_0$

This algorithm requires $[\log_2 n] + v(n)$ multiplications, where $v(n)$ is the number of ones in the binary representation of n. A similar algorithm can be written for the polynomial exponentiation of Algorithms 1 and 2.

(c) *Back conversion* : Let each prime in the prime coding scheme require at most b bits for representation, and let $M_n(k)$ be the time required to multiply two k-bit numbers. An $O(M_n(br) \log r)$ preconditioned algorithm for back conversion is available in [1]. A similar algorithm for polynomial moduli is given below.

Algorithm 4. *Back conversion*

Input

(a) Relatively prime polynomial moduli over $GF(p)$, $\phi_1(x), \phi_2(x), \ldots, \phi_r(x)$, where $r = 2^t$ for some t.

(b) A set of inverses $d_1(x), d_2(x), \ldots, d_r(x)$ such that $d_i(x) = [\phi(x)/\phi_i(x)]^{-1} \bmod \phi_i(x)$, where

$$\phi(x) = \prod_{i=1}^{r} \phi_i(x).$$

(c) A sequence of residues $w_1(x), w_2(x), \ldots, w_r(x)$.

Output. The unique polynomial $M(x)$ over $GF(p)$ with degree less than $\phi(x)$, satisfying the congruences $w_i(x) = M(x) \bmod \phi_i(x)$, $i = 1, 2, \ldots, r$.

Algorithm

begin$_0$

1. *for* $i \leftarrow 1$ *until* r *do* $S_{i0}(x) \leftarrow d_i(x) \cdot w_i(x)$

2. *for* $j \leftarrow 1$ *until* t *do*

begin$_1$

3. *for* $i \leftarrow 1$ *step* 2^j *until* r *do*

begin$_2$

4. $S_{ij}(x) \leftarrow S_{i,j-1}(x) \cdot q_{i+2^{j-1},j-1}(x) + S_{i+2^{j-1},j-1}(x) \cdot q_{i,j-1}(x)$

Comment $q_{ij}(x) = \prod_{m=i}^{i+2^j-1} \phi_m(x)$

end$_2$

end$_1$

5. *write* $S_{1t}(x) \bmod q_{1t}(x)$

end$_0$

If each $\phi_i(x)$ is of degree d, then the complexity of this algorithm is $O(dr \log r \log dr)$.

(d) *GCD and congruence inverse* : The standard method of computing the *gcd* of two integers a and b is Euclid's algorithm. Euclid's algorithm can be extended to find integer multipliers x and y such that

$$xa + yb = (a, b).$$

This algorithm is available in [8] and [1]. This algorithm can be used to generate the inverse of a mod b when $(a, b) = 1$. In this case we have

$$xa + yb = 1$$

i.e., $$xa = 1 \bmod b.$$

Thus the premultiplier x is the required inverse of a mod b.

The *gcd* algorithm is required at the key generation stage to ensure that the chosen n satisfies the criterion $(n, p_i - 1) = 1$, $i = 1, 2, \ldots, r$ for primes (or $(n, p^{d_i} - 1) = 1$ for irreducible polynomials). Similarly, the inverse algorithm is required to generate the m_i's for the exponent at the decoding end. These two algorithms are required only when the key is changed, and hence their complexity does not affect the complexity of the encoding-decoding process.

(e) *Choice of moduli.* The choice of primes is determined largely by the criterion

$$P = \prod_{i=1}^{r} p_i > 2^{mk}.$$

As r becomes larger, the decoding operation becomes more efficient, since the b_i's and m_i's become smaller. The effect of increasing r on the number of keys is not very clear, and will probably depend on the structure of the $(p_i - 1)$, $i = 1, 2, \ldots, r$. The number P should be chosen close to the bound to keep the size of the transmitted code as small as possible.

In the polynomial case, the overall set of irreducible polynomials is determined once p, r_i's and d_i's are fixed. A method for generating such polynomials is given in [3], but the procedure is very complicated. Tables of such polynomials have been constructed, and it is advisable to refer to them where feasible. Alanen and Knuth [2] for instance, tabulate all indexing polynomials for fields containing 1024 elements or less. (An indexing polynomial is an irreducible polynomial, all of whose roots lie in the field generated by it).

As r increases, the decoding operation becomes more efficient in the polynomial case also. The effect on the number of keys is again not clear, since $(p^d - 1)$ decreases with increasing r while the behaviour of

$$\binom{I_{d,p}}{r}$$

is not easy to predict ($I_{d,p}$ is expected to decrease with increasing r). However, it is clear that having irreducible polynomials of different degrees increases the number of keys (for a fixed m), since this choice increases $Q(\phi)$ and hence $\Phi(Q(\phi))$.

(f) *Comparison of prime and polynomial systems* : In comparing the prime and polynomial systems, the following point becomes clear. In the prime system, the cryptanalyst can form a fairly good idea of the value of P as he intercepts more of the transmitted message. For example, let $P = 4199$. As mentioned earlier, the cryptanalyst knows the system being used, i.e., he knows the block size. The message intercepted by him will contain no block whose value is greater than $P = 4199$, while the values occurring in the message will give him a lower bound for P which will come closer to P as more of the message is intercepted. Once the cryptanalyst determines P from this method, the values of the p_i can be determined by factoring P.

In the polynomial case, however, each irreducible polynomial generates the *same* finite field. Hence, different sets of moduli will generate elements from the same set, and hence, the cryptanalyst can obtain no further information about the key by studying the transmitted blocks. Decoding by studying the pattern of occurrence of the blocks can be made impractical by making the block-length large. The cryptanalyst's best method of attack will be to run through the possible combinations of n and the $\phi_i(x)$ to determine which combination gives a meaningful decoded message. The key should be changed sufficiently often to prevent deciphering by such a method (that is, the probability of such a deciphering taking place should be made vanishingly small). Since the number of elements in the key is small, the key can be changed without much inconvenience.

In view of this, a cipher system using irreducible polynomials appears more attractive than a similar system using primes, especially when the degrees of the irreducible polynomials are different.

(g) *Complexity* : Let $M_n(b)$ be the complexity of multiplying two b-bit numbers and let $M_p(k)$ be the complexity of multiplying two kth degree polynomials. Since the encoding and decoding operations are similar, we will consider only the complexity of the decoding operation. We assume that all irreducible polynomials have the same degree d. The decoding operation for polynomials requires

$$O(r \log m_0 \cdot M_p(d) + \log r \cdot M(dr)) \text{ time,}$$

where $m_1 = m_2 = \ldots = m_r = m_0$. The first term represents the time required to execute step 7 of Algorithm 2. The second term gives the time required for combining r terms by Chinese Remaindering. (The calculation of the residues in step 6 of Algorithm 2 requires an equivalent time [1]). The decoding operation for primes requires

$$O\left[\sum_{i=1}^{r} \log m_i \cdot M_n(b_i) + \log r M_n(b)\right],$$

time, where b_i is the number of bits required to represent p_i and $b = \sum_{i=1}^{r} b_i$.

Cryptographic systems involving finite field transforms have been suggested in connection with public-key cryptography [10]. Rivest's system is similar to our prime coding system and its secrecy depends on the complexity of factoring the product of two very large primes. Our system, on the other hand, has been developed for conventional cryptographic use, in which the entire key is assumed to be inaccessible to the cryptanalyst and the key is changed periodically for security. In our system, the primes can be much smaller than those used by Rivest and this obviously reduces the complexity of encoding and decoding processes. In fact, under these circumstances, we have already shown (in the previous point (f)) that the use of irreducible polynomials is superior to the use of primes. Rivest's system cannot be readily extended to the polynomial case since there are polynomial-time algorithms for factoring a polynomial over a finite field [4], [5].

We should also like to mention here that our method of coding and decoding is more efficient than Rivest's method, which directly determines M^n mod P for encoding (and E^{m_0} mod P for decoding). To compare the complexities of the two methods, let us assume that b bits are required to represent P and b' bits are required to represent each p_i, $i = 1, 2, \ldots, r$. If the individual p_i's are of the same order of magnitude, then $b \simeq rb'$. We further assume that $n = \prod_{i=1}^{r} n_i$ (in practice n can be greater than, equal to or less than $\prod_{i=1}^{r} n_i$). Since the value of n_i can range from 1 to $\Phi(p_i)$, n requires, on the average, $b'/2$ bits for representation, i.e., $O(b')$ bits. Rivest's method for encoding requires $O(M_n(b) \cdot \log n) = t_1$ time while our method for primes requires $O(M_n(b') \cdot \log n_i + M_n(b) \cdot \log r) = t_2$ time, i.e.,

$$t_1 = O(b \cdot M_n(b)),$$

$$t_2 = O(b' \cdot M_n(b') + \log(r) \cdot M_n(b)).$$

Depending on which term dominates in t_2, we obtain the ratio t_1/t_2 as either $b \cdot M_n(b)/M_n(b') \cdot b'$ or $b/\log r$. Assuming $M_n(b)$ to be a linear function of b, we obtain $b \cdot M_n(b)/M_n(b') \cdot b' = r^2$. More reasonable assumptions for $M_n(b)$ such as $b \cdot \log b \cdot \log \log b$ or b^k, $1 < k \leqslant 2$, give a higher value for this ratio. Thus our method is clearly more efficient than Rivest's method. A similar consideration applies for the decoding procedure. The reason for this clear improvement in efficiency is not merely the use of modular arithmetic, but the fact that the exponent for each residue is much smaller than n. Hence, the number of multiplications for each residue is decreased, in addition to the decrease in size.

7. Conclusion

The congruence cipher presented in this paper allows the use of block-coding with an extremely small size for the associated key. This cipher will be very useful in a number of cases where pre-processing to remove the statistics of the message source is inconvenient or undesirable. An 8-character block-coding scheme using the English alphabet set can be constructed using 4 different irreducible polynomials of degree 2 over $GF(29)$. This system will have around 10^{11} different keys and an inexpensive microprocessor-based system can be easily developed to implement this scheme. Each key in this system contains only 9 elements (the value of n, and the coefficients of x and the constant term for the 4 irreducible polynomials). Hence the key can be changed often without inconvenience. The number of different keys can be increased by using polynomials of different degrees. It is expected that similar systems will find wide application in military communications and computer data-protection.

References

[1] Aho A V, Hopcropt J E and Ullman J D 1974 *The Design and analysis of computer algorithms.* (Reading, MA : Addison Wesley)
[2] Alanen J D and Knuth D E 1964 *Sankhya (Calcutta)* **A26** 305
[3] Albert A A 1956 *Fundamental concepts of higher algebra* (Chicago : University Press)
[4] Berlekamp E R 1967 *Bell Syst. Tech. J.* **46** 1853
[5] Berlekamp E R 1968 *Algebraic coding theory* (New York : McGraw-Hill)
[6] Berlekamp E R 1970 *Math. Computation* **24** 713
[7] Diffie W and Hellman M E 1976 *IEEE Trans. Inform. Theory* **22** 644
[8] Knuth D E 1969 *The art of computer programming, Vol.* 2, *Semi-numerical algorithms* (Reading, MA : Addison Wesely)
[9] Niven I and Zuckerman H S 1960 *An introduction to the theory of numbers* (New York : John Wiley)
[10] Rivest R L, Shamir A and Adleman L 1978 *Communications of the Association for Computing Machinery* **21** 120
[11] Shannon C E 1949 *Bell Syst. Tech. J.* **28** 656

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 95–100.

Measurability of inverses of random operators and existence theorems

MOHAN JOSHI
Department of Mathematics, Birla Institute of Technology and Science, Pilani 333 031, India

MS received 29 December 1978; revised 9 June 1979

Abstract. Let $(\Omega, \mathcal{B}, \mu)$ be ameasure space and X a separable Hilbert space. Let T be a random operator from $\Omega \times X$ into X. In this paper we investigate the measurability of T^{-1}. In our main theorems we show that if T is a separable random operator with $T(\omega)$ almost sure invertible and monotone and demi-continuous then T^{-1} is also a random operator. As an application of this we give an existence theorem for random Hammerstein operator equation.

Keywords. Separable random operator; monotone operator; Hammerstein operator; existence theorems.

1. Introduction

Nashed and Salehi [5] obtained the following theorem on the measurability of the inverse of random non-linear operator.

Theorem 1.1

Let $(\Omega, \mathcal{B}, \mu)$ be a complete probability space, X be a separable metric space and Y be a metric space. Let T be a separable random operator from $\Omega \times X$ onto Y such that almost sure $T(\omega)$ is invertible and its inverse $T^{-1}(\omega)$ is continuous. Then T^{-1} is also a random operator from $\Omega \times Y$ into X.

In our main theorem in this paper we obtain a similar result on a Hilbert space $\mathcal{X}$ without the condition of continuity on the inverse operator $T^{-1}(\omega)$. Instead, we impose the monotonicity condition on the operator $T(\omega)$. The statement of our main theorem reads as follows.

Theorem 1.2

Let $(\Omega, \mathcal{B}, \mu)$ be a complete probability space and $\mathcal{X}$ be a separable Hilbert space. Let T be a separable random operator from $\Omega \times \mathcal{X}$ onto $\mathcal{X}$ such that almost sure $T(\omega)$ is invertible and monotone and demi-continuous. Then T^{-1} is also a random operator from $\Omega \times \mathcal{X}$ into $\mathcal{X}$.

This theorem is then followed by an important result regarding the solvability of random Hammerstein equation with an application to concrete random non-linear integral equation.

2. Preliminaries

Let $(\Omega, \mathfrak{B}, \mu)$ be a probability space with a probability measure μ; that is, Ω is a nonempty set, $\mathfrak{B}$ is the σ-algebra of subsets of Ω and μ is a probability measure. We say that the probability space is complete if $B \in \mathfrak{B}$, $\mu(B) = 0$ and $B_0 \subseteq B$ mplies that $B_0 \in \mathfrak{B}$.

A function g from Ω into a normed space Y is Y-valued random variable if the inverse image, under the function g, of each Borel set $B \in \mathfrak{B}_Y$ belongs to $\mathfrak{B}$ where $\mathfrak{B}_Y$ is the σ-algebra generated by closed subsets of Y.

The mapping T from $\Omega \times \Gamma$ into Y, where Γ an arbitrary set is called a random operator if for each $\gamma \in \Gamma$, the function $T(\cdot)\,\gamma$ is a random variable.

A random operator $T(\omega) : X \to Y$ said to be continuous at $x_0 \in X$ if $x_n \to x_0$ implies that $T(\omega)\,x_n \to T(\omega)\,x_0$ almost surely. It is called demi-continuous if convergence of $T(\omega)\,x_n$ to $T(\omega)\,x_0$ is weak.

Theorem 2.1

Let g be a random variable with values in a separable Banach space X and let T be a continuous random operator of the space $\Omega \times X$ into metric space Z. Then the mapping W of Ω into Z defined by, for every $\omega \in \Omega$, $W(\omega) = T(\omega)\,g(\omega)$ is a random variable with values in Z.

A random operator T from $\Omega \times X$ into Y, where Ω is a complete probability space, X a separable metric space and Y a metric space, is said to be separable if there exists a countable set $S \subset X$ and negligible set $N \in \mathfrak{B}$, $\mu(N) = 0$, such that

$$\{\omega : T(\omega)\,x \in K;\ \ x \in F \cap S\} \triangle \{\omega : T(\omega)\,x \in K, x \in F\} \subset N$$

for every closed set K in $\mathfrak{B}_Y$ and every F in $\mathfrak{B}_X$.

For a further study of separable random operators we refer to [2]. It is easy to see that the above definition of separability is equivalent to the following: there exists a negligible set $N \in \mathfrak{B}$ and a countable set $S \subset X$ such that for $\omega \notin N$ and each $x \in X$ there exists a sequence $\{x_i\} \in S$ such that $x_i \to x$ and $T(\omega)\,x_i \to T(\omega)\,x$. We can now state the following result [1].

Theorem 2.2

Let X be a separable Banach space and $T : \Omega \times X \to X$ be a continuous random operator. Then T is separable.

Let T be a random operator from $\Omega \times X$ into Y. An equation of the type $T(\cdot)\,x(\cdot) = y(\cdot)$ where y is a given random variable with values in Y is called a random operator equation. Any X-valued random variable $x(\omega)$ which satisfies

$$\mu\{\omega : T(\omega)\,x(\omega) = y(\omega)\} = 1$$

is said to be random solution of the above equation.

We now give few important definitions and theorems regarding monotone operators. In what follows X is a Banach space $\langle , \rangle$ a bilinear form on $X \times X^*$ and T a nonlinear operator from X into X^*.

T is called monotone if $\langle Tx_1 - Tx_2, x_1 - x_2 \rangle \geqslant 0$ for all $x_1, x_2 \in X$. T is called strictly monotone if the above inequality is strict for $x_1 \neq x_2$. T is strongly mono-

tone if there exists $c > 0$ such that $\langle Tx_1 - Tx_2, x_1 - x_2\rangle \geqslant c \| x_1 - x_2 \|^2$ for all x_1, x_2 in X.

We have the following theorems regarding monotone operators. For reference see [3].

Theorem 2.3

Let T be a demi-continuous monotone operator from a Banach space X to its dual X^* such that

$$\langle y - Tx, x_0 - x\rangle \geqslant 0 \text{ for all } x \in X, \text{ then } y = Tx_0.$$

Theorem 2.4

Let T be a demi-continuous strongly monotone operator from X into X^*. Then T is 1-1 and onto with T^{-1} continuous.

We say that $T : X \to X^*$ is coercive if $\langle Tx, x\rangle / \| x \| \to \infty$ as $\| x \| \to \infty$.

Theorem 2.5

Let $T : X \to X^*$ be a demi-continuous, monotone and coercive operator. Then $R(T) = X^*$.

Finally we give a definition, called angle-boundedness, for the bounded monotone linear operator K from X to X^*. A bounded monotone linear operator K from X to X^* is called angle bounded if there exists a constant $\alpha \geqslant 0$ such that

$$| \langle Kx, y\rangle - \langle Ky, x\rangle | \leqslant 2\alpha [\langle Kx, x\rangle \langle Ky, y\rangle]^{1/2}$$

for x, y in X.

It is clear that every symmetric, monotone linear operator is angle-bounded with constant zero.

3. Existence theorems

Let X be a separable Hilbert space and $(\Omega, \mathfrak{B}, \mu)$ a complete probability measure space. $T : \Omega \times X \to X$ is a random operator. Following is the main theorem of this paper.

Theorem 3.1

Let T be a separable random operator from $\Omega \times X$ onto X such that almost sure $T(\omega)$ is invertible and monotone and demi-continuous. Then T^{-1} is also a random operator from $\Omega \times X$ into X.

Proof

Let superscript c denote the complementation, $S(., r)$ and $\bar{S}(.,r)$ denote the open and closed ball of radius r around respectively. To prove the measurability of T^{-1} it suffices to show that for an arbitrary $y \in X$ and a closed ball $\bar{S}(x, r)$ the event $\{\omega : T^{-1}(\omega)\, y \in \bar{S}(x', r)\}$ is in B. We have

$$\{\omega : T^{-1}(\omega)\, y \in \bar{S}(x', r)\} = \bigcup_{s \in \bar{S}(x', r)} \{\omega : T(\omega)\, x = y\}$$

We claim that

$$\bigcup_{x \in \bar{S}(x', r)} \{\omega : T(\omega)\, x = y\} = \bigcap_{n=1}^{\infty} \bigcup_{x \in S(x', r + 1/n)} \{\omega : T(\omega)\, x \in S(y, 1/n)\}. \quad (3.1)$$

It is enough to show that the right side is contained in the left side.
Let

$$\omega_0 \in \bigcap_{n=1}^{\infty} \bigcup_{x \in S(x', r + 1/n)} \{\omega : T(\omega)\, x \in S(y, 1/n)\}.$$

Then for each n there exists $x_n \in S(x', r + 1/n)$ such that $T(\omega_0)\, x_n \in S(y, 1/n)$. It follows that

$$\lim_{n \to \infty} T(\omega_0)\, x_n = y.$$

Since the sequence $\{x_n\}$ is a bounded sequence in a Hilbert space, it follows that there exists a subsequence

$$\{x_{n_k}\},\ x_{n_k} \in S\left(x',\ r + \frac{1}{n_k}\right),$$

and x_0 such that $x_{n_k} \to x_0$ weakly. We claim that $x_0 \in \bar{S}(x', r)$. This follows easily from the relation

$$\begin{aligned} \langle x_0 - x', x_0 - x' \rangle &= \langle x_0 - x_{n_k}, x_0 - x' \rangle + \langle x_{n_k} - x', x_0 - x' \rangle \\ &\leqslant | \langle x_0 - x_{n_k}, x_0 - x' \rangle | + || x_{n_k} - x' ||\ ||\, x_0 - x' ||, \end{aligned}$$

using the fact that $x_{n_k} \to x_0$ weakly and $x_{n_k} \in S\left(x', r + \frac{1}{n_k}\right)$.

Moreover, we have

$$\langle T(\omega_0)\, x_{n_k} - T(\omega_0)\, x, x_{n_k} - x \rangle \geqslant 0 \text{ for all } x \in X.$$

Now, since

$$\lim_{k \to \infty} T(\omega_0)\, x_{n_k} = y \text{ and } x_{n_k} \to x_0$$

weakly; passing over to the limit we get

$$\langle y - T(\omega_0)\, x, x_0 - x \rangle \geqslant 0 \text{ for all } x \in X.$$

Since $T(\omega_0)$ is monotone and demi-continuous, it follows by theorem 2.3 that $y = T(\omega_0)\, x_0$. This together with the fact that

$$x_0 \in \bar{S}(x', r) \text{ implies that } \omega_0 \in \bigcup_{\bar{S}(x', r)} \{\omega : T(\omega)\, x = y\}.$$

But

$$\left[\bigcap_{n=1}^{\infty} \bigcup_{S\left(x, r + \frac{1}{n}\right)} \left\{\omega :\ T(\omega)\, x \in S\left(y, \frac{1}{n}\right)\right\}\right]^c$$

$$= \bigcup_{n=1}^{\infty} \bigcap_{S\left(x', r + \frac{1}{n}\right)} \left\{\omega : T(\omega)\, x \in S^c\left(y, \frac{1}{n}\right)\right\} \quad (3.2)$$

Because of separability of T,

$$\bigcap_{s\left(x', r+\frac{1}{n}\right)} \left\{\omega : T(\omega)\, x \in S^e\left(y, \frac{1}{n}\right)\right\}$$

is measurable. This together with (3.1) and (3.2) gives the result.

As a corollary we can get existence and uniqueness of a random solution $x(\omega)$ of the operator equation $T(\omega)\, x = y(\omega)$.

Corollary 3.1

Let $T : \Omega \times X \to X$ be a continuous random operator such that almost surely $T(\omega)$ is strictly monotone and coercive. Then there exists a unique random solution $x(\omega)$ of the operator equation $T(\omega)\, x(\omega) = y(\omega)$.

Proof. Since $T(\omega)$ is strictly monotone and coercive it follows from the theorem 2.5 that $T(\omega)$ is 1–1 and onto and hence almost sure $T(\omega)$ is invertible. Further, since X is separable and T is continuous, it follows by theorem 2.2 that T is separable. Thus T satisfies all the conditions of the above theorem and hence T^{-1} is also a random operator from $\Omega \times X$ into X. That is, there exists a random variable $x(\omega)$ such that $T(\omega)\, x(\omega) = y(\omega)$. Uniqueness of $x(\omega)$ follows from 1-1 property of $T(\omega)$.

As an application we now state an existence result for random Hammerstein operator equation

$$x(\omega) + KNx(\omega) = y(\omega)$$

on a separable Banach space X. Here $K : \Omega \times X \to X^*$ is a random linear operator and $N : \Omega \times X^* \to X$ is a random nonlinear operator.

Theorem 3.2

Let X be a separable Banach space and let

(i) $K(\omega) : X \to X^*$ be a continuous random monotone operator with $\| K(\omega) \| \leqslant K_0$ and with a fixed constant of angle boundedness α.

(ii) $N(\omega) : X^* \to X$ be a continuous random operator such that
$\langle x - y, N(\omega)\, x - N(\omega)\, y \rangle \geqslant -k(\omega) \| x - y \|^2$ for all $x, y \in X^*$ (3.3)
and for almost all $\omega \in \Omega$.

Suppose that $k(\omega)\, K_0 (1 + \alpha^2) < 1$ for almost all $\omega \in \Omega$, then there exists a unique random solution $x(\omega)$ in X^* such that

$$x(\omega) + KNx(\omega) = y(\omega). \tag{3.4}$$

This theorem generalises the Browder and Gupta's theorem [3] to random Hammerstein equations. The proof is similar to that of Browder [3].

4. Example

We now give an example which depicts the application of theorem 3.2 to a concrete nonlinear integral equation. We consider an equation of the type

$$x(s; \omega) + \int_{\Sigma} K(s, t; \omega) f(t, x(t); \omega)\, dt = y(s; \omega) \tag{4.1}$$

where (i) $\omega \in \Omega$, Ω is the supporting set of the probability space $(\Omega, \mathcal{B}, \mu)$,

(ii) Σ is a σ-finite measure space,

(iii) $K(s, t ; \omega)$ is a random kernel defined on $\Sigma \times \Sigma$,

(iv) $f(t, x, \omega)$ is a nonlinear random function defined on $\Sigma \times R^n$ with values in R^n,

(v) $y(s ; \omega)$ is a known and $x(s ; \omega)$ is an unknown n-dimensional valued random variable defined for $s \in \Sigma$.

In order to consider this equation for the existence of a random solution we transfer it into random Hammerstein operator equation. We define the random linear operator K and the random nonlinear operator N as

$$K(\omega)\, x(s) = \int_{\Sigma} K(s, t ; \omega)\, x(t)\, dt$$

$$N(\omega)\, x(t) = f(t, x(t) ; \omega).$$

(4.1) is then equivalent to the random operator equation

$$x(\omega) + KNx(\omega) = y(\omega). \tag{4.2}$$

We assume that the function $f(t, u ; \omega)$ satisfy the conditions

(i) $|f(t, x ; \omega)| \leqslant b(\omega)\,[g(t) + b\,|x|]$, $g \in L^2(\Sigma)$, $b(\omega) > 0$.

(ii) $(f(t, x ; \omega) - f(t, y ; \omega))(x - y) \geqslant 0$ for almost all $\omega \in \Omega$.

The random kernel $K(s, t ; \omega)$ is assumed to belong to $L^2(\Sigma \times \Sigma)$ with $||K(s, t ; \omega)|| \leqslant K_0$, for every $\omega \in \Omega$. Further we assume that it is symmetric and monotone for almost all $\omega \in \Omega$. Under these conditions it can be easily seen that the operator $K(\omega)$ is a continuous random linear angle bounded operator with constant $\alpha = 0$. The nonlinear operator $N(\omega)$ is also continuous and monotone. Hence it follows by theorem 3.2 that there exists a unique random solution $x(\omega)$ of (4.1).

References

[1] Bharucha-Reid A T 1972 *Random integral equations* (New York : Academic Press)

[2] Bharucha-Reid A T and Mukherjea A 1969 *Rev. Roumaine Math. Pures Appl.* **14** 1553

[3] Browder F E 1971 *Contributions to nonlinear functional analysis* ed. E Zarantonello (New York : Academic Press) p. 99

[4] Minty G J 1963 *Proc. Natl. Acad. Sci. USA* **50** 1038

[5] Nashed M Z and Salehi H 1973 *SIAM J. Appl. Math.* **25** 681

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 101–102.

On Sharma-Swarup algorithm for time minimising transportation problems

C R SESHAN and V G TIKEKAR
Department of Applied Mathematics, Indian Institute of Science, Bangalore 560 012, India
MS received 11 January 1979; revised 11 July 1979

Abstract. In this note, the fallacy in the method given by Sharma and Swarup, in their paper on time minimising transportation problem, to determine the set S_{hk} of all nonbasic cells which when introduced into the basis, either would eliminate a given basic cell (h, k) from the basis or reduce the amount x_{hk} is pointed out.

Keywords. Time minimising transportation problem; basic cells; nonbasic cells.

Recently, an algorithm has been given by Sharma and Swarup [6] for solving the problem of transportation in minimum time. The algorithm is very similar to the one given by Hammer [3], [4] and the only place where it differs from Hammer's algorithm is in the determination of the set S_{hk} defined as follows (see step 2.2 of [6]):

$$S_{hk} = \{(i,j) \mid (i,j) \text{ is a nonbasic cell which, when introduced into the basis, either would eliminate the cell } (h, k) \text{ from the basis or reduce the amount } x_{hk}\}$$

Here (h, k) is a predetermined basic cell in a given basic feasible solution.

The method proposed in [6] to determine the set S_{hk} is incorrect and when applied to the numerical example given by them, fails to determine actual S_{hk}. For the basic feasible solution X_1 (see [6]) it is easy to see that $(h, k) = (3, 5)$ and actual S_{hk} is given by

$$S_{hk} = \{(1, 5), (2, 5), (1, 6), (2, 6), (3, 6)\}.$$

Further the set $\{(2, 5), (1, 6), (2, 6), (3, 6)\} \subset S_{hk}$ is the set of nonbasic cells (i, j) such that if (i, j) is introduced into the basis, the resultant adjacent basic feasible solution will be better than the current one. If we implement the method given by Sharma and Swarup [6], we shall get

$$S_{hk} = \{(2, 1), (3, 1), (4, 1), (2, 2), (3, 2), (4, 2), (3, 3), (4, 3), (1, 4), (4, 4), (1, 5), (2, 5), (1, 6), (2, 6), (3, 6)\}$$

which is not correct.

The fallacy is due to the fact that while the determination of S_{hk} has nothing to do with the actual values of t_{ij}, and depends only on the set of basic cells, the method given by Sharma and Swarup [6] is based on the actual values of t_{ij}. It may be noted that the definition of the set S_{hk} given in step 2.2 and by the last line of p. 516 [see [6]) are inconsistent.

The correct method of finding S_{hk} (as in Hammer [3]) is briefly stated below.

Divide the set M of basic cells into levels $L_0, L_1, L_2, \cdots$ as follows. The level L_0 consists of only the cell (h, k). The level L_{p+1} $(p \geq 0)$ consists of all those basic cells which are in the same row or column as a basic cell in level L_p and which are not in level L_q for any $q \leq p$. Let $p(i)$ $(q(j))$ be the least index p such that there exists a basic cell in level L_p in row i (column j). Define

$$u(i) = 0 \quad \text{if } p(i) \text{ is even}$$
$$= 1 \text{ if } p(i) \text{ is odd}$$
$$v(j) = 0 \quad \text{if } q(j) \text{ is even}$$
$$= 1 \quad \text{if } q(j) \text{ is odd}$$
$$S_{hk} = \{(i,j) \mid (i,j) \text{ is a nonbasic cell and } u(i) = v(j) = 0\}.$$

Note that $u(i)$, $v(j)$ can also be determined by observing that they are the dual variables for the following costs of the basic cells:

$$C_{hk} = 0, \quad C_{ij} = 1 \text{ if } (i,j) \text{ is basic and } (i,j) \neq (h,k)$$

obtained by setting $u_h = 0$.

It may also be noted that Szwarc [7], Garfinkel and Rao [2], Ramakrishnan [5] and Bhatia *et al* [1] have studied the time minimising transportation problem, the algorithm proposed by Garfinkel and Rao [2] being known as 'threshold algorithm'.

Acknowledgement

The authors wish to thank the referee for his useful comments.

References

[1] Bhatia H L, Kanti Swarup and Puri M C 1977 *Indian J. Pure Appl. Math.* **8** 920
[2] Garfinkel R S and Rao M R 1971 *Nav. Res. Log. Q.* **18** 465
[3] Hammer P L 1969 *Nav. Res. Log. Q.* **16** 345
[4] Hammer P L 1971 *Nav. Res. Log. Q.* **18** 487
[5] Ramakrishnan C S 1977 *Opsearch* **14** 207
[6] Sharma J K and Kanti Swarup 1977 *Proc. Indian Acad. Sci. (Math. Sci.)* **A86** 513
[7] Szwarc W 1971 *Nav. Res. Log. Q.* **18** 473

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 103–112.

MHD Couette flow of a viscous stratified fluid of large conductivity

K N VENKATASIVA MURTHY
Department of Mathematics, Post-graduate Centre, Anantapur 515 003, India

MS received 31 May 1977; revised 30 July 1979

Abstract. The MHD Couette flow of a viscous stratified fluid of large electrical conductivity with suction and injection at the plane boundaries is studied when the plane boundaries are maintained at different temperatures. The Oseen type governing equations are formulated using the method suggested by Greenspan for stratified fluids. Introducing the similarity variables, the linearised equations are solved to obtain the velocity and temperature distributions. The results show that the behaviour of velocity and temperature in fluids of large conductivity is different from the behaviour of velocity and temperature for fluids of finite conductivity. The effect of the magnetic field on the load capacity is investigated for the case when the width of the channel is small.

Keywords. Suction; injection; stratification; load capacity.

1. Introduction

One of the basic problems, the plane Couette flow, has been a source for investigation of many research workers in dealing with the interplay of various fluid forces and their interaction with the electromagnetic forces. Sinha and Choudhary [6], Terril and Sreshta [8], Verma and Bansal [11] and Verma and Bhatt [10] used this simple configuration to study the effect of suction and injection at the plane boundaries, which is an important device in controlling the phenomena of seperation, etc. This model is also used by Leadon [3] and Bleviss [1] to investigate the effect of heat transfer on the fluid flow by maintaining the plane boundaries at equal or different temperatures. The effect of uniform magnetic field on the Couette flow is investigated by Yen and Chang [12] and Soundalgekar [7]. The Couette flow results are also useful in the theory of hydrodynamic lubrication. In view of these applications to modern technology, the Couette flow has attracted the attention of several research workers.

In dealing with the MHD problems, it is sometimes useful to make the assumption of infinite electrical conductivity to obtain quantitative information about physical situations [9, 4, 5]. This assumption generally allows a simple mathematical formulation and it is of interest to investigate how the results of the Couette flow of a fluid of large electrical conductivity differ from the results already obtained for the

Couette flow of a fluid of finite conductivity. Also, the results of the Couette flow are useful in designing the slider bearing systems. It will be of importance to find the effect of the magnetic field on the load capacity when the lubricant is of large electrical conductivity. Motivated by this, we investigate the MHD Couette flow of a fluid of large electrical conductivity with suction and injection at the plane boundaries maintained at different temperatures. The uniform magnetic field is applied transverse to the electrically insulating plane boundaries. The Oseen type governing equations are obtained using the method suggested by Greenspan [2] for stratified fluids. The effect of suction and injection, the stratification and the magnetic field on the velocity and temperature distributions is discussed. The load per unit length which the bearing can support is calculated and its response to the change in the magnetic field is studied. It is shown that the results obtained for a fluid of large conductivity remarkably differ from the corresponding results for a fluid of finite conductivity, for some ranges of values of the magnetic parameter.

2. Governing equations and solution

We consider the motion of a viscous, stratified, conducting fluid between two infinite parallel planes which are maintained at different temperatures. The lower plane is assumed to be stationary and the upper plane moves with a uniform velocity $\epsilon U'$ in its plane. The fluid is subjected to uniform injection at the upper plane and uniform suction at the lower plane. The system is under the influence of a uniform magnetic field transverse to the plane electrically insulating boundaries. We make the following assumptions:

(i) The fluid is of large electrical conductivity.

(ii) The influence of density variation with temperature is considered only on the body force term in accordance with the Boussinesq approximation.

(iii) Viscous dissipation and Joule heating are neglected.

(iv) All physical properties of the fluid like viscosity, electrical conductivity and magnetic permeability are constant. The MHD equations of steady motion are

$$\rho' \bar{q}' \cdot \nabla \bar{q}' = - \nabla p' + \mu \nabla^2 \bar{q}' - \rho' g \bar{k} + \mu_e \bar{J} \times \bar{H},$$

$$\nabla \cdot \bar{q}' = 0,$$

$$\rho' c_p \bar{q}' \cdot \nabla T' = k \nabla^2 T',$$

$$\bar{J} = \nabla \times \bar{H},$$

$$(\bar{H} \cdot \nabla) \bar{q}' = (\bar{q}' \cdot \nabla) \bar{H},$$

$$\rho' - \rho_e = - \alpha (T' - T_e),$$

$$\rho_e = \rho_0 - \alpha (T_1 - T_2) z'/L,$$

$$T_e = T_0 + (T_1 - T_2) z'/L,$$

where $2L$ is the distance between the plane boundaries located at $z' = \pm L$, $\bar{q}' = (u', 0, - W' + w')$ is the velocity vector, $\bar{k}$ the unit vector in the z' direction, ρ',

p', T', $\bar{J}$ are the density, pressure, temperature, current density, $\bar{H} = (h_x', O, H_0 + h_z')$ is the magnetic field, μ_e the magnetic permeability, μ the coefficient of viscosity and α the stratification parameter. ρ_e, T_e and p_e are the equilibrium values T_1 and T_2 are the temperatures at the upper and lower planes and $T_0 = (T_1 + T_2)/2$. The boundary conditions are

$$u' = \epsilon U', \; w' = -W'\epsilon, \; T' = T_1, \text{ when } z' = L/2,$$
$$u' = 0, \; w' = W'\epsilon, \; T' = T_2, \text{ when } z' = -L/2.$$

We introduce the following non-dimensional variables:

$$u' = \epsilon\beta u, \; w' = \epsilon\beta w, \; x' = xL, \; z' = zL, \; h_x' = \epsilon H_0 h_x, \; h_z' = \epsilon H_0 h_z,$$
$$\rho' = \rho_e + \alpha\epsilon(T_1 - T_2)\rho,$$
$$p' = \rho_0 g L p_e + \epsilon\alpha g L (T_1 - T_2) p,$$
$$T' = T_e + \epsilon(T_1 - T_2)\theta,$$

where ϵ is a small non-dimensional parameter and

$$\beta = [\alpha(T_1 - T_2) gL/\rho_0]^{1/2}.$$

With these substitutions, the equations governing the steady motion in the non-dimensional form are

$$(1 - Hz + \epsilon H\rho)\left(\epsilon u\frac{\partial u}{\partial x} + \epsilon w\frac{\partial u}{\partial z} - \delta\frac{\partial u}{\partial z}\right)$$
$$= -\frac{\partial p}{\partial x} + E\nabla^2 u + \frac{\mu_e H_0^2}{\rho_0\beta^2}\left(\frac{\partial h_x}{\partial z} - \frac{\partial h_z}{\partial x}\right)(1 + \epsilon h_z),$$
$$(1 - Hz + \epsilon H\rho)\left(\epsilon u\frac{\partial w}{\partial x} + \epsilon w\frac{\partial w}{\partial z} - \delta\frac{\partial w}{\partial z}\right)$$
$$= -\frac{\partial p}{\partial z} + E\nabla^2 w + T - \frac{\mu_e H_0^2}{\rho_0\beta^2}\epsilon h_x\left(\frac{\partial h_x}{\partial z} - \frac{\partial h_z}{\partial x}\right),$$
$$(1 - Hz + \epsilon H\rho)\left(\epsilon u\frac{\partial\theta}{\partial x} + \epsilon w\frac{\partial\theta}{\partial z} + w - \delta\frac{\partial\theta}{\partial z} - \frac{\delta}{\epsilon}\right) = \frac{E}{P}\nabla^2\theta,$$
$$\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z} = 0,$$
$$\epsilon u\frac{\partial h_x}{\partial x} + (\epsilon w - \delta)\frac{\partial h_x}{\partial z} = \epsilon h_x\frac{\partial u}{\partial x} + (1 + \epsilon h_z)\frac{\partial u}{\partial z},$$
$$\epsilon u\frac{\partial h_z}{\partial x} + (\epsilon w - \delta)\frac{\partial h_z}{\partial z} = \epsilon h_x\frac{\partial w}{\partial x} + (1 + \epsilon h_z)\frac{\partial w}{\partial z},$$

where $H = \alpha(T_1 - T_2)/\rho_0$ is the density ratio, $P = \mu c_p/k$ is the Prandtl number, $E = \mu/\rho_0\,\beta L$ and $\delta = W'/\beta$. Since h and w are uniform on the plane boundaries, we assume $w = w(z)$, $h_x = h_x(z)$. Assuming ϵ and H to be small and neglecting terms corresponding to these small parameters, the above equations reduce to

$$-\delta\frac{\partial u}{\partial z} = -\frac{\partial p}{\partial x} + E\nabla^2 u - M\frac{\partial u}{\partial z}, \tag{1}$$

$$-\delta \frac{\partial w}{\partial z} = -\frac{\partial p}{\partial z} + E \nabla^2 w + \theta, \tag{2}$$

$$-\delta \frac{\partial \theta}{\partial z} = \frac{E}{P} \nabla^2 \theta - w + \frac{\delta}{\epsilon}, \tag{3}$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial z^2}$ and $M = \mu_e H_0^2 / W' \beta \rho_0$ is the magnetic parameter. In view of the equation of continuity, we define the similarity variables

$$w = -f(z), \; u = xf'(z) + g(z)$$

where f and g are unknown functions to be determined. Eliminating p from (1) and (2) and substituting for u and w in terms of f and g we obtain

$$\frac{\partial \theta}{\partial x} = x(Ef^{iv} + \delta g''' - Mf''') + (Eg''' + \delta g'' - Mg'') \tag{4}$$

where the prime denotes differentiation with respect to z. From this equation, it follows that the appropriate form for θ is

$$\theta(x, z) = x^2 \theta_2(z) + x\theta_1(z) + \theta_0(z).$$

Substituting for θ in equation (3) we obtain the differential equations governing θ_0, θ_1, and θ_2 as

$$(E/P)\theta_i'' + \delta\theta_i = 0 \quad (i = 1, 2), \tag{5}$$

$$(E/P)\theta_i'' + \delta\theta_0' + (2E/P)\theta_2 + f + (\delta/\epsilon) = 0. \tag{6}$$

The boundary conditions on θ_0, θ_1 and θ_2 take the form

$$\theta_0 = \theta_1 = \theta_2 = 0 \text{ on } z = \pm 1/2.$$

Solving (5) with the above boundary conditions, we obtain $\theta_1 = \theta_2 = 0$ and hence $\theta = \theta_0(z)$. Substituting for θ in (2) we find that $(\partial p/\partial z)$ is a function of z alone. Equation (1) then determines the appropriate form for p as

$$p(x, z) = F(z) + Ax^2 + Bx + C,$$

where A, B and C are constants. Choosing the origin such that $B = 0$, $A \neq 0$, it follows that

$$f^{iv} + (\delta/E)f''' - (M/E)f''' = 0, \tag{7}$$

$$g'' + (\delta/E)g' - (M/E)g' = 0. \tag{8}$$

The boundary conditions in terms of f and g are

$$f = \delta, \; f' = 0, \; g = U, \text{ on } z = \tfrac{1}{2},$$

$$f = -\delta, \; f' = 0, \; g = 0, \text{ on } z = -\tfrac{1}{2},$$

where $U' = \beta U$. The general solutions for f and g satisfying the required boundary conditions are

$$f = A + Bz + Cz^2 + D \exp(2Nz),$$

$$g = F + G \exp(2Nz),$$

where $N = (M - \delta)/2E$,

$A = \delta\,[-(N/2)\ \sinh N + \cosh\ N]/\phi\,(N)$,

$B = 2\delta N \cosh\ N/\phi\,(N)$,

$C = B\ \tanh\ N$,

$D = -\delta/\phi\,(N)$,

$F = -U\ \exp(-N)/2\ \sinh\ N$,

$G = U/2\ \sinh\ N$,

$\phi\,(N) = N \cosh\ N - \sinh\ N$.

Solving equations (6) for $\theta_0(z)$ with the boundary conditions $\theta_0 = 0$ on $z = \pm \frac{1}{2}$, we obtain

$$\theta_0 = Z_1 + Z_2\ \exp\ (-P\delta z/E) + \psi\,(z)$$

where $$\psi\,(z) = (-C/3\delta)\,z^3 + \left(\frac{CE}{\delta^2 P} - \frac{B}{2\delta}\right)z^2 + \left(\frac{EB}{\delta^2 P} - \frac{2E^2 C}{\delta^3 P^2} - \frac{A}{\delta} - \frac{1}{\epsilon}\right)z$$

$$- \frac{PD}{4EN^2 + 2NP\delta}\ \exp\ (2Nz),$$

$$Z_1 = -Z_2 \exp\,(-P\delta/2E) - \psi\,(\tfrac{1}{2}),$$

$$Z_2 = \frac{\psi\,(\frac{1}{2}) - \psi\,(-\frac{1}{2})}{\sinh\,(P\delta/2E)}.$$

When $\delta = M$, the solution is given by

$f = \delta\,(3z - 4z^3)$, $g = 0{\cdot}5 + z$ and

$$\theta_0(z) = z^4 - \frac{4E}{P\delta}z^3 + \left(\frac{12E^2}{\delta^2 P^2} - 1{\cdot}5\right)z^2 + \left(\frac{3E}{P\delta} - \frac{1}{\epsilon} - \frac{24E^3}{\delta^2 P^3}\right)z.$$

3. Discussion of the results

The velocity and temperature profiles are drawn in figures 1 to 4. The non-dimensional velocity u is drawn with z in figure 1 for various values of the magnetic parameter M. In the case of an MHD Couette flow of a fluid of finite conductivity it was observed by Soundalgekar [7] and also by Yen and Chang [12] that the velocity decreases with the increase in the Hartmann number. In contrast to this, for a fluid of large conductivity, we obtain (figure 1) that if $M > \delta$, the velocity increases as M increases at distances near the upper plane whereas for distances near the lower plane the velocity decreases as M increases. There is a steep rise in the velocity in a layer near the upper plane and the thickness of the layer decreases as M increases. This boundary layer-type behaviour of the velocity profile can also be observed by considering the solution for values of M such that $(M - \delta)/E \gg 1$. For such values of M we have

$$u = 2\epsilon Wx\,(1 + 2z - 2\ \exp\,[(M - \delta)\,(2z - 1)/2E] + U\ \exp\,[(M - \delta)$$

$$(2z - 1)/2E].$$

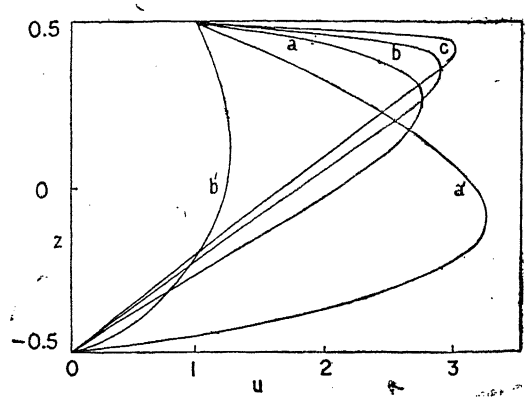

Figure 1. Horizontal velocity when $E = 0{\cdot}008$, $\delta = 0{\cdot}04$, $U = 1$, $x = 20$.
(a) $M = 0{\cdot}1$ (b) $M = 0{\cdot}4$ (c) $M = 0{\cdot}6$ (a′) $M = 0{\cdot}01$ (b′) $M = 0{\cdot}02$.

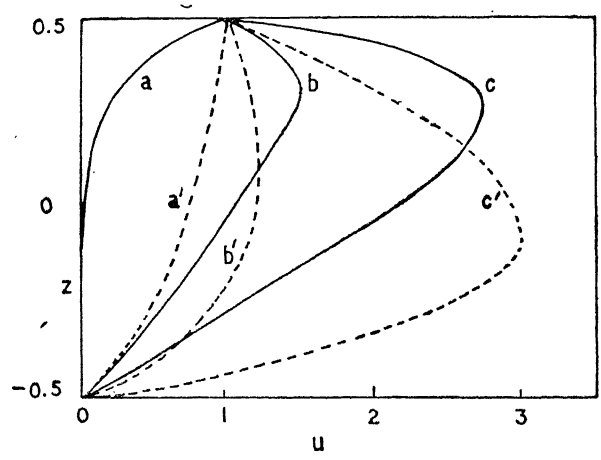

Figure 2. Horizontal velocity when $E = 0{\cdot}008$, $\delta = 0{\cdot}04$, $U = 1$.
——— $M > \delta$ - - - - $M < \delta$
(a) $x = 0$ (b) $x = 10$ (c) $x = 20$ (M=0·2)
(a′) $x = 0$ (b′) $x = 20$ (c′) $x = 100$ (M=0·02)

The exponential terms in this solution represent boundary layer near the upper plane of thickness order $E/(M - \delta)$ and the remaining part $2\epsilon Wx(1 + 2z)$ represents the flow outside this layer. Thus, for sufficiently large values of M except in a boundary layer of thickness order $E/(M - \delta)$ near the upper plane, the distribution is linear. When $M < \delta$, the horizontal velocity decreases as M increases.

The horizontal velocity at various cross-sections of the channel is shown in figure 2. It can be observed that the horizontal velocity increases as x increases. The maximum velocity occurs at distances near the upper plane for $M > \delta$ and at distances near the lower plane for $M < \delta$.

In the non-magnetic case, it was observed by Verma and Bansal [11] that the flow from large values of x to the mouth of the channel is developed near the statio-

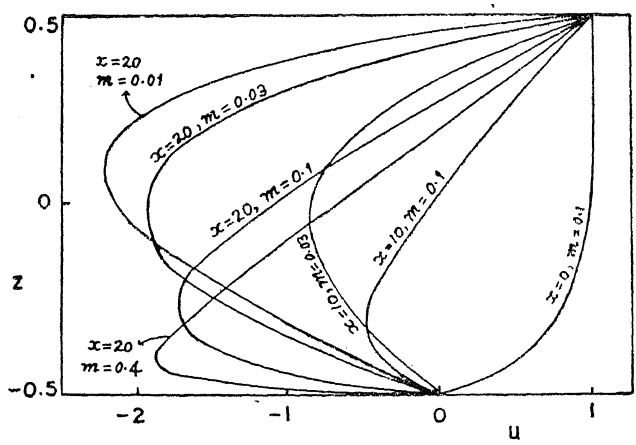

Figure 3. Velocity distribution when W' is negative, $d = 0{\cdot}04$, $E = 0{\cdot}008$.

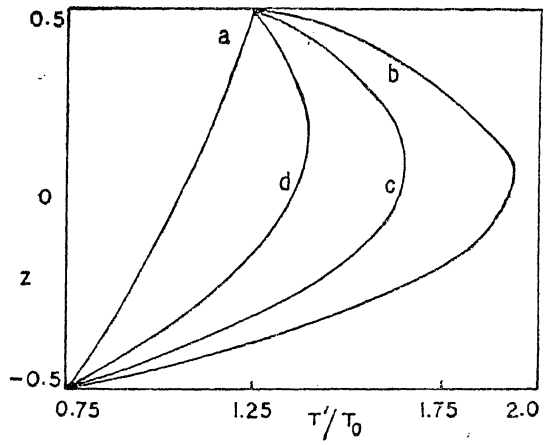

Figure 4. Temperature distribution when $\delta = E = 1$, $\epsilon = 0{\cdot}01$
(a) $M = 0{\cdot}2$, $P = 0{\cdot}4$ (b) $M = 0{\cdot}9$, $P = 0{\cdot}4$
(c) $M = 0{\cdot}9$, $P = 0{\cdot}6$ (d) $M = 0{\cdot}9$, $P = 0{\cdot}8$.

nary wall. Thus the dragging action of the faster layers exerted on the fluid particles in the neighbourhood of the stationary wall is insufficient to overcome the influence of the adverse pressure gradient that develops. The problem described by Verma and Bansal is comparable to the present problem if the suction velocity W' is negative. The behaviour of the flow for negative values of W' is described by figure 3. The velocity u is plotted with z for various values of the magnetic parameter $m = \mu_e H_0^2/(-W')\beta\rho_0$ and the suction parameter $d = -W'/\beta$. The behaviour of the flow near the lower plane in the case of negative W' is similar to that near the upper plane in the case of positive W'.

The temperature distribution is shown in figure 4. When $M < \delta$, the temperature increases as M increases. For values of M nearer to δ, there is a steep rise in the temperature near the lower plane. When $M > \delta$, the temperature decreases as M increases (table 1).

The effect of stratification on the temperature distribution can be seen by changing the values of M, δ and E such that $(M - \delta)/E$ remains the same (since M, δ and E depend on α and $(M - \delta)/E$ is independent of α). The results are tabulated in table 2. It can be observed that for any value of the magnetic parameter M, a stronger stratification produces an increase in the temperature.

Table 1 shows that as the difference in the suction velocities at the two planes increases, the temperature also increases.

3.1. *Load capacity*

When the width of the channel is small, the Couette flow results can be used to calculate the load which the bearing can support. From equations (1) and (2), the appropriate form for p' can be obtained as

$$p' = \rho_0 g L P_e + \epsilon \alpha g L (T_1 - T_2) p,$$

where $$p = (\delta - M) Cx^2 + Mf''(z) + \theta_0(z) + K,$$

Table 1. $E = \delta = 1, P = 0{\cdot}4.$

M	ϵ	$z=-0{\cdot}5$	$z=-0{\cdot}3$	$z=-0{\cdot}1$	$z=0{\cdot}1$	$z=0{\cdot}3$	$z=0{\cdot}5$
0·9	0·005	0·75	1·162	1·410	1·503	1·447	1·25
	0·01	0·75	1·457	1·845	1·931	1·729	1·25
0·2	0·005	0·75	0·871	0·981	1·081	1·171	1·25
	0·01	0·75	0·876	0·988	1·089	1·176	1·25
1·5	0·005	0·75	0·880	0·992	1·090	1·176	1·25
	0·02	0·75	0·919	1·046	1·140	1·206	1·25
3·0	0·005	0·75	0·868	0·976	1·075	1·166	1·25
	0·02	0·75	0·872	0·981	1·079	1·168	1·25

Table 2. $(M - \delta)/E = 0{\cdot}1,\ P = 0{\cdot}4,\ \epsilon = 0{\cdot}01.$

z Value		$-0{\cdot}5$	$-0{\cdot}3$	$-0{\cdot}1$	$0{\cdot}1$	$0{\cdot}3$	$0{\cdot}5$
$M = 0{\cdot}9, \delta = E = 1{\cdot}0$		0·75	1·456	1·845	1·931	1·729	1·25
0·45	0·5	0·75	2·047	2·717	2·789	2·792	1·25
0·2	1·0	0·75	0·876	0·988	1·089	1·176	1·25
0·1	0·5	0·75	0·884	1·003	1·104	1·187	1·25
1·5	1·0	0·75	0·893	1·011	1·107	1·186	1·25
0·75	0·5	0·75	0·919	1·047	1·140	1·206	1·25
3·0	1·0	0·75	0·869	0·977	1·076	1·167	1·25
1·5	0·5	0·75	0·872	0·981	1·079	1·168	1·25

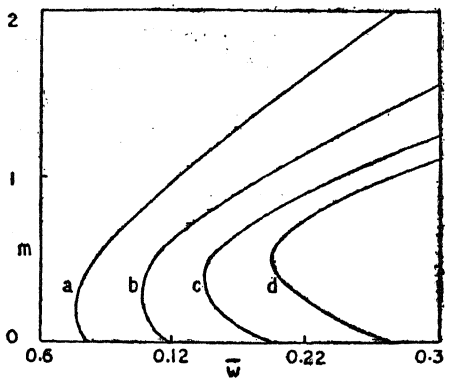

Figure 5. Load capacity plotted with the magnetic parameter m when $E = 0{\cdot}1$. (a) $d = 0{\cdot}2$ (b) $d = 0{\cdot}3$ (c) $d = 0{\cdot}4$ (d) $d = 0{\cdot}5$.

where K is a constant. Since the width of the channel is small the pressure $p(x, z)$ is assumed constant, $p(x, -\frac{1}{2})$ in the thin film. The constant K can be found by matching the pressure at the throat section $x = 0$ of the channel to the pressure p_0 outside,

$$p = p_0 \text{ at } x = 0.$$

The non-dimensional load per unit length which the bearing can support is defined by

$$\bar{W} = \int_0^1 (\bar{p}' - \bar{p}_0)\, dx,$$

where $\bar{p}' = p'/\epsilon\alpha g L (T_1 - T_2)$,

and $\bar{p}_0 = p_0/\epsilon\alpha g L (T_1 - T_2)$.

The dependence of the load capacity on the parameters m and d is shown in figure 5. For any given value of d, the load capacity first decreases and then increases as m increases from zero. This is in contrast to the case of a fluid of finite conductivity where the load capacity increases with the strength of the magnetic field. Also, as the parameter d increases the load capacity decreases.

Acknowledgement

The author is thankful to the referee for critical comments enabling improvement in the paper.

References

[1] Bleviss Z D 1958 *J. Aero Sci.* **25** 601
[2] Greenspan H P 1969 *The theory of rotating fluids* (Cambridge: University Press) p. 15
[3] Leadon B M 1957 Convair Sci. Res. Lab. RN 13

[4] Sarma L V K V 1962a *Appl. Sci. Res.* **B9** 245
[5] Sarma L V K V 1962b *Sci. and Eng.* Third annual number
[6] Sinha K D and Choudhary R C 1965 *Proc. Indian Acad. Sci.* **A61** 308
[7] Soundalgekar V M 1966 *Proc. Indian Acad. Sci.* **A64** 304
[8] Terril R M and Sreshta G M 1965 *Z. Angew. Math. Phys.* **16** 470
[9] Venkatasiva Murthy K N and Seetharamaswamy R 1978 *Proc. Indian Acad. Sci.* (*Math. Sci.*) **A87** 247
[10] Verma P D and Bhatt B S 1973 *Indian J. Phys.* **47** 718
[11] Verma P D and Bansal J L 1966 *Proc. Indian Acad. Sci.* **A64** 385
[12] Yen J T and Chang C C 1964 *Z. Angew. Math. Phys.* **15** 400

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 113–123.

Forced convection over a semi-infinite flat plate

N K BANTHIYA and NOOR AFZAL*
Department of Mechanical Engineering, Technical Teachers Training Institute, Bhopal 462 003, India
* Department of Mechanical Engineering, Aligarh Muslim University, Aligarh 202 001, India

MS received 14 April 1979

Abstract. The problem of forced convection heat transfer over a semi-infinite flat plate is treated by the method of series truncation, so as to yield results valid from leading edge to far downstream ($0 \leqslant R_x < \infty$). Results are presented for Prandtl number Pr = 0·1, 0·7, and 10. It is found that the effect of leading edge on heat transfer is smaller than on skin friction.

Keywords. Boundary layer; heat transfer.

List of Symbols

c_{fx}	local skin friction coefficient $[\tau_{wx}/(\rho U_\infty^2/2)]$
c_p	specific heat at constant pressure
$f(\xi, \eta)$	non-dimensional stream function (ψ/ν)
h_x	local heat transfer coefficient
i	$\sqrt{-1}$
k	fluid thermal conductivity
Nu_x	local Nusselt number $(h_x x/k)$
Pr	Prandtl number $(c_p \mu/k)$
R_x	local Reynolds number $(U_\infty x/\nu)$
T	temperature
U_∞	free stream velocity
x	co-ordinate measured along the plate
y	co-ordinate measured normal to the plate

Greek symbols

δ_x	viscous layer thickness

δ_t thermal layer thickness

δ^* displacement thickness

$\triangle T_{ref}$ reference temperature difference

η non-dimensional parabolic co-ordinate

$\theta(\xi, \eta)$ non-dimensional temperature difference $(T - T_\infty)/\triangle T_{ref}$

$\bar{\theta}(\xi)$ non-dimensional temperature difference in equation (10)

$\theta_w(\xi_0)$ non-dimensional temperature difference at the wall

$\theta_1, \theta_2, \theta_3$ first, second and third order temperature difference in equation (10)

μ dynamic viscosity of fluid

ν kinematic viscosity of fluid

ξ non-dimensional parabolic co-ordinate

ξ_0 $(2R_x)^{1/2}$

τ_{wx} local wall shear stress

ψ stream function

ψ_1, ψ_2, ψ_3 first, second and third order stream functions in equation (9).

Subscripts

w Wall

∞ Far away from the wall.

1. Introduction

In the present paper the problem of forced convection heat transfer over a semi-infinite flat plate with arbitrary surface temperature variation is formulated; so as to yield results valid from leading edge to far downstream ($0 \leqslant R_x < \infty$), R_x being the Reynolds number based on streamwise distance x. The method of local series truncation Davis [2] is employed and numerical results are presented for an isothermal plate and Prandtl numbers Pr = 0·1, 0·7, and 10.

Although a large number of solutions for momentum equation for small or moderate values of R_x have been attempted (to be discussed), no such solutions for the corresponding thermal problem have been reported; except the solution of Van Dyke [2] based on second-order boundary layer theory, which for a semi-infinite flat plate does not yield any correction to the classical boundary-layer solution. A brief review of the solutions of momentum equation now follows.

Carrier and Lin [1] obtained a solution for momentum equation valid near the leading edge, based on Stokes flow. The region of validity of this solution is quite small. This solution could not be properly matched and shows the appearance

of several undetermined constants. An interesting analysis by Imai [6], based on global momentum balance in a large contour gives the second term in the expression for integrated skin friction. The validity of boundary-layer solution has been extended for moderately large values of R_x by obtaining higher order approximations by Goldstein [5], Van Dyke [11, 12], and Murray [8]. However, these solutions cannot yield correctly the integrated skin friction just by integrating, due to nonintegrable singularity at the leading edge. These solutions also contain eigen functions with undetermined multiplying constants, whose values depend on the details of the flow near the leading edge. Additional solutions by Yoshizawa [15] and Van de Vooren and Dijkstra [9] based on integration of full Navier-Stokes equations have become available recently. Dean [4] and Davis [2] give solutions to Navier-Stokes equations valid from leading edge to far downstream. Dean uses the boundary-layer solution as a first approximation to evaluate the nonlinear convective terms in the Navier-Stokes equations, to construct solution near the leading edge. Davis employs the local similarity approach *via* the semi-analytical method of series truncation, reporting results which agree very well with those of Dean and Imai.

The method of series truncation has been successfully employed for a number of problems in fluid mechanics, see Van Dyke [13, 14]. It has been found especially suitable for elliptic partial differential equations, as it reduces these to ordinary differential equations; which are simpler to solve than the original partial differential equations. This method employs series expansions for the dependent variables in terms of one of the independent variables; whereas functional dependence on the other independent variables is left unspecified. The system of ordinary differential equations obtained by substituting these series in governing partial differential equations display upstream influence, due to ellipticity of the problem and hence it is not closed. However, by truncating the series at any desired stage the system can be closed. This summary truncation assuming higher order excess unknowns to be zero though somewhat brutal, yields extremely good results; if the co-ordinate system and form of expansions are chosen judiciously see e.g. Van Dyke [13]. A variant of series truncation is the method of local series truncation, employed by Davis [2], in which, instead of representing the entire flow field by a single expansion from the same point, a number of expansions about different co-ordinate lines are constructed and the results of such solutions are used only locally.

2. Governing equations

Consider a semi-infinite flat plate with zero angle of attack. The rectangular cartesian co-ordinate x is along the plate, measured from the leading edge and y normal to it measured from the wall. The non-dimensional parabolic co-ordinates (ξ, η) employed are defined as (figure 1),

$$x + iy = \nu(\xi + i\eta)^2/2U_\infty, \tag{1}$$

$$x = \nu(\xi^2 - \eta^2)/2U_\infty, \quad y = \nu\xi\eta/U_\infty. \tag{2}$$

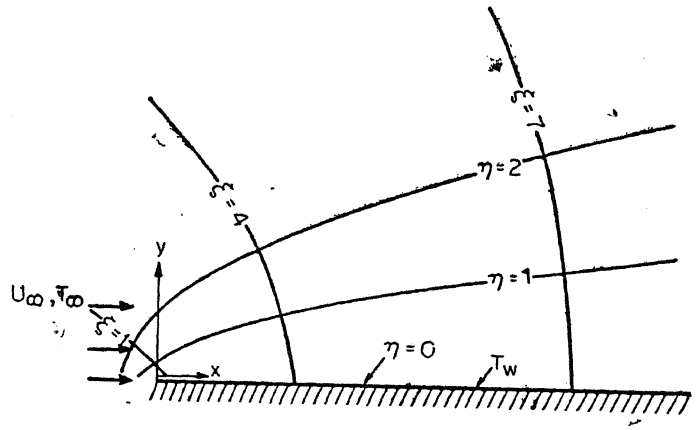

Figure 1. Co-ordinate system.

Parabolic co-ordinates have been employed being optimal [7]. The value of ξ at the wall denoted by ξ_0 is given by

$$\xi_0 = (2R_x)^{1/2} \tag{3}$$

where R_x is the local value of Reynolds number $(U_\infty x/\nu)$.

The governing equations for steady, laminar, two-dimensional, incompressible flow, neglecting viscous dissipation in (ξ, η) co-ordinates are

$$\text{Vorticity}: \quad \left[\left(\frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2}\right) + f_\xi \left(\frac{\partial}{\partial \eta}\right) - f_\eta \left(\frac{\partial}{\partial \xi}\right)\right]\left[\frac{f_{\xi\xi} + f_{\eta\eta}}{\xi^2 + \eta^2}\right] = 0, \tag{4}$$

$$\text{Energy}: \quad \theta_{\xi\xi} + \theta_{\eta\eta} + \Pr\,(f_\xi \theta_\eta - f_\eta \theta_\xi) = 0. \tag{5}$$

The corresponding boundary conditions are

$$\eta = 0; \; f = 0, \; \partial f/\partial \eta = 0, \qquad \theta = \theta_w(\xi_0), \tag{6}$$

$$\eta \to \infty; \; f \sim \xi\eta, \; \theta \to 0. \tag{7}$$

Wall temperature is considered to be prescribed in an arbitrary manner as

$$(T_w - T_\infty) = (\triangle T_{\text{ref.}})\, \theta_w(\xi_0). \tag{8}$$

3. Application of local series truncation to governing equations

In the present work, the scheme of local series truncation is employed. Stream function f and temperature difference θ are expanded about an arbitrary point ξ_0 on the plate as (see also Davis [2]),

$$f = \xi[\psi_1(\eta) + (\xi - \xi_0)\psi_2(\eta) + (\xi - \xi_0)^2 \psi_3(\eta) + \cdots], \tag{9}$$

$$\theta = \bar{\theta}(\xi)\,[\theta_1(\eta) + (\xi - \xi_0)\,\theta_2(\eta) + (\xi - \xi_0)^2\,\theta_3(\eta) + \cdots]. \tag{10}$$

The form of expansions in (9) and (10) are primarily dictated by the boundary conditions. In (10), $\bar{\theta}(\xi)$ is a function of ξ only, its form depending on prescribed wall temperature distribution. By substituting only the first two terms of (9) and (10) in the governing equations, following equations for second truncation, valid

about an arbitrary point ξ_0 on the plate are obtained.

Vorticity: Coefficient of $(\xi - \xi_0)^0$

$$\psi_1'''' + \left[\psi_1 + \xi_0\psi_2 - \frac{4\eta}{\xi_0^2+\eta^2}\right]\psi_1''' + \frac{1}{\xi_0^2+\eta^2}\left[(\xi_0^2-\eta^2)\psi_1' - 2\eta\psi_1 - 2\xi_0\eta\psi_2\right]\psi_1'' + \left[-\xi_0\psi_2'' + \frac{4\xi_0}{\xi_0^2+\eta^2}\psi_2\right]\psi_1' + 2\left[\frac{\psi_2'}{\xi_0} - \frac{2\eta}{\xi_0(\xi_0^2+\eta^2)}\psi_2\right]\psi_1 + \frac{4\eta^2}{\xi_0(\xi_0^2+\eta^2)}\psi_2'' + 2\left[\psi_2 - \frac{4\eta}{\xi_0(\xi_0^2+\eta^2)}\right]\psi_2' + \frac{4}{\xi_0^2+\eta^2}\left[-\eta\psi_2 + \frac{2}{\xi_0}\right]\psi_2 = 0, \quad (11)$$

Coefficient of $(\xi - \xi_0)$

$$\psi_2'''' + \left[\psi_1 + \xi_0\psi_2 - \frac{4\eta}{\xi_0^2+\eta^2}\right]\psi_2''' - \left[3\psi_1' + \frac{2\eta}{\xi_0^2+\eta^2}\psi_1 + \xi_0\psi_2' + \frac{2\eta\xi_0}{\xi_0^2+\eta^2}\psi_2\right]\psi_2'' + \frac{1}{\xi_0^2+\eta^2}\left[(\xi_0^2-\eta^2)\psi_1'' + 4\psi_1 + \frac{4(3\xi_0^2+\eta^2)}{\xi_0}\psi_2\right]\psi_2' + \frac{1}{\xi_0^2+\eta^2}\left[(5\xi_0^2+3\eta^2)\psi_1''' - 6\eta\psi_1'' + 8\psi_1' - \frac{8\eta}{\xi_0}\psi_2\right]\psi_2 + \frac{3\xi_0^2+\eta^2}{\xi_0(\xi_0^2+\eta^2)}\psi_1'''' + \frac{1}{\xi_0(\xi_0^2+\eta^2)} \times \left[(3\xi_0^2+\eta^2)\psi_1 - 4\eta\right]\psi_1''' + \frac{1}{\xi_0(\xi_0^2+\eta^2)} \times \left[(3\xi_0^2-\eta^2)\psi_1' - 2\eta\psi_1\right]\psi_1'' = 0, \quad (12)$$

Energy: Coefficient of $(\xi - \xi_0)^0$

$$\theta_1'' + \Pr(\psi_1 + \xi_0\psi_2)\theta_1' + \left[-\Pr\xi_0\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\psi_1' + \left(\frac{d^2\bar{\theta}}{d\xi^2}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right]\theta_1 + \left[-\Pr\xi_0\psi_1' + 2\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right]\theta_2 = 0, \quad (13)$$

Coefficient of $(\xi - \xi_0)$

$$\theta_2'' + \Pr(\psi_1 + \xi_0\psi_2)\theta_2' + \left[-\Pr\left\{1 + 2\xi_0\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right\}\psi_1' - \Pr\xi_0\psi_2' + 3\left(\frac{d^2\bar{\theta}}{d\xi^2}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right]\theta_2 + \left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\theta_1'' + \Pr\left[(\psi_1 + \xi_0\psi_2) \times \left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0} + 2\psi_2\right]\theta_1' + \left[-\Pr\left\{\xi_0\left(\frac{d^2\bar{\theta}}{d\xi^2}\Big/\bar{\theta}\right)_{\xi=\xi_0} + \left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right\}\psi_1' - \Pr\xi_0\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\psi_2' + \left(\frac{d^3\bar{\theta}}{d\xi^3}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right]\theta_1 = 0. \quad (14)$$

The associated boundary conditions are

$$\eta = 0; \quad \psi_1 = \psi_1' = 0; \quad \psi_2 = \psi_2' = 0; \tag{15}$$

$$\theta_1 = 1, \quad \theta_2 = 0; \tag{16}$$

$$\eta \to \infty; \ \psi_1' \to 1, \psi_2' \to 0; \tag{17}$$

$$\theta_1 \to 0, \quad \theta_2 \to 0. \tag{18}$$

Equations (11) and (12) are similar to those given by Davis, except that there are some errors in his equations. To yield correct equations, the term ψ_2, in Davis' equations is to be replaced by ψ_2/ξ_0. For an isothermal plate energy equations (13) and (14) reduce to

$$\theta_1'' + \Pr\ (\psi_1 + \xi_0\psi_2)\,\theta_1' - \Pr\ \xi_0\psi_1'\,\theta_2 = 0, \tag{19}$$

$$\theta_2'' + \Pr\ (\psi_1 + \xi_0\psi_2)\,\theta_2' - \Pr\ [\psi_1' + \xi_0\psi_2']\,\theta_2 + 2\Pr\,\psi_2\theta_1' = 0. \tag{20}$$

The above equations are valid at all points on the plate, except at the leading edge $\xi_0 = 0$. To study the behaviour near the leading edge expansions can be centred around the leading edge itself. The series for f and θ can be assumed as

$$f = \xi\,[\psi_1(\eta) + \xi^2\psi_3(\eta) + \cdots], \tag{21}$$

$$\theta = \bar{\theta}(\xi)\ \ [\theta_1(\eta) + \xi^2\,\theta_3(\eta) + \cdots]. \tag{22}$$

The later terms in the above expansions do not follow in simple powers of ξ, see Carrier and Lin [1], and Davis [2].

It is not possible to obtain second truncation energy equations valid at the leading edge for any arbitrary wall temperature distribution. Some of the wall conditions which are amenable to analysis are isothermal wall, linear variation of temperature and exponential variation of temperature.

First truncation equations, for an arbitrary wall temperature distribution and valid about an arbitrary point ξ_0 are

$$\text{Vorticity:}\quad \psi_1'''' + \left[\psi_1 - \frac{4\eta}{\xi_0^2 + \eta^2}\right]\psi_1''' + \frac{1}{\xi_0^2 + \eta^2}\left[(\xi_0^2 - \eta^2)\,\psi_1' - 2\eta\psi_1\right]\psi_1'' = 0, \tag{23}$$

$$\text{Energy:}\quad \theta_1'' + \Pr\ \psi_1\theta_1' + \left[-\Pr\ \xi_0\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)_{\xi=\xi_0}\psi_1' + \left(\frac{d^2\bar{\theta}}{d\xi^2}\Big/\bar{\theta}\right)_{\xi=\xi_0}\right]\theta_1 = 0. \tag{24}$$

The associated boundary conditions are

$$\eta = 0, \ \psi_1 = \psi_1' = 0, \tag{25}$$

$$\theta_1 = 1, \tag{26}$$

$$\eta \to \infty, \ \psi_1' \to 1, \ \psi_1'' \to 0, \tag{27}$$

$$\theta_1 \to 0. \tag{28}$$

First truncation equations valid at the leading edge obtained by expansions (21) and (22) or by (23) and (24) taking $\xi_0 = 0$ turn out to be same and are

$$\text{Vorticity:}\quad \psi_1'''' + \left[\psi_1 - \frac{4}{\eta}\right]\psi_1'' - \left[\psi_1' + \frac{2}{\eta}\psi_1\right]\psi_1'' = 0, \tag{29}$$

$$\text{Energy:}\quad \theta_1'' + \text{Pr}\ \psi_1\theta_1' + \left[-\text{Pr}\ \xi\left(\frac{d\bar{\theta}}{d\xi}\Big/\bar{\theta}\right)\psi_1' + \left(\frac{d^2\bar{\theta}}{d\xi^2}\Big/\bar{\theta}\right)\right]_{\xi=0}\theta_1 = 0. \tag{30}$$

If the variation of $\bar{\theta}(\xi)$ is specified equation (30) can be solved if no singularity exists in the equation at $\xi = 0$, e.g. for an isothermal wall (30) simplifies to

$$\theta_1'' + \text{Pr}\ \psi_1\theta_1' = 0. \tag{31}$$

4. Skin friction and heat transfer

Non-dimensional skin friction $[c_{fx}(R_x/2)^{1/2}]$ and non-dimensional heat transfer $[\text{Nu}_x/(R_x/2)^{1/2}]$, when the assumed series expansions are considered will be given in both first and second truncations by

$$c_{fx}(R_x/2)^{1/2} = \psi_1''(0), \tag{32}$$

$$\text{Nu}_x/(R_x/2)^{1/2} = -\bar{\theta}(\xi_0)\theta_1'(0). \tag{33}$$

5. Numerical method of solution

For integrating second truncation equations (11) to (14) as an initial value problem by Runge-Kutta method four initial conditions for the vorticity and two for the energy equation will have to be guessed. This guesswork can be reduced by one for the vorticity equation, if a group property of the equations can be found which eliminates ξ_0 from the equation, the transformation being

$$\eta \to \xi_0\eta,\quad \psi_1 \to (\psi_1/\xi_0),\quad \psi_2 \to (\psi_2/\xi_0^2). \tag{34}$$

However, no general transformation exists for θ_1 and θ_2 which may eliminate ξ_0 from the energy equation; although it is possible to find transformations for specific wall temperature distributions. For an isothermal wall the transformation is

$$\theta_1 \to \theta_1,\ \theta_2 \to (\theta_2/\xi_0). \tag{35}$$

The numerical solution can now be obtained by guessing three initial values for the vorticity equation and two for the energy equation. In this solution, the Reynolds number cannot be fixed *a priori*, it comes as a part of the solution when the infinity condition $\psi_1''(\infty) \sim 0$ is satisfied, as $\psi_1'(\infty) \sim \xi_0^2$.

The solution of first truncation equations (23)–(24) after the transformation (34)–(35) is much simpler as only one initial value for the vorticity equation and one for the energy equation will have to be guessed. In the present work the first truncation results for a large number of values of R_x were evaluated for isothermal wall condition and Pr = 0·1, 0·7 and 10. When the second truncation solutions were obtained for a few representative values of R_x at Pr = 0·7, no appreciable

change in the value of Nu_x, as compared to first truncation values was found. Hence attention was concentrated on first truncation solutions. The numerical results were obtained by incorporating Newton's method of interpolation to obtain newer estimates of starting values for numerical integration. The numerical integration was done using fourth order Runge-Kutta method with Gill's modification, and using double precision arithmetic on IBM 7044/1401 system at Indian Institute of Technology, Kanpur.

For obtaining solutions valid at the leading edge, only first truncation equations (29) and (31) have been integrated. The solution of equation (29) has been obtained in the manner discussed by Davis [2]. The solution of equation (31) is now straightforward.

6. Results and discussion

Velocity and temperature profiles displayed in figures 2 and 3 show that around $R_x = 1000$ boundary-layer Blasius profiles are obtained. At lower values of R_x, the profiles become more fuller than boundary-layer profiles. Further, the thickness of viscous layer $\delta_s(\eta)$ and thermal layer $\delta_t(\eta)$ also increase as R_x increases to approach the Blasius values. In the physical plane it implies that at large R_x when $\delta_s(\eta)$ approaches a constant value the displacement thickness δ^* grows

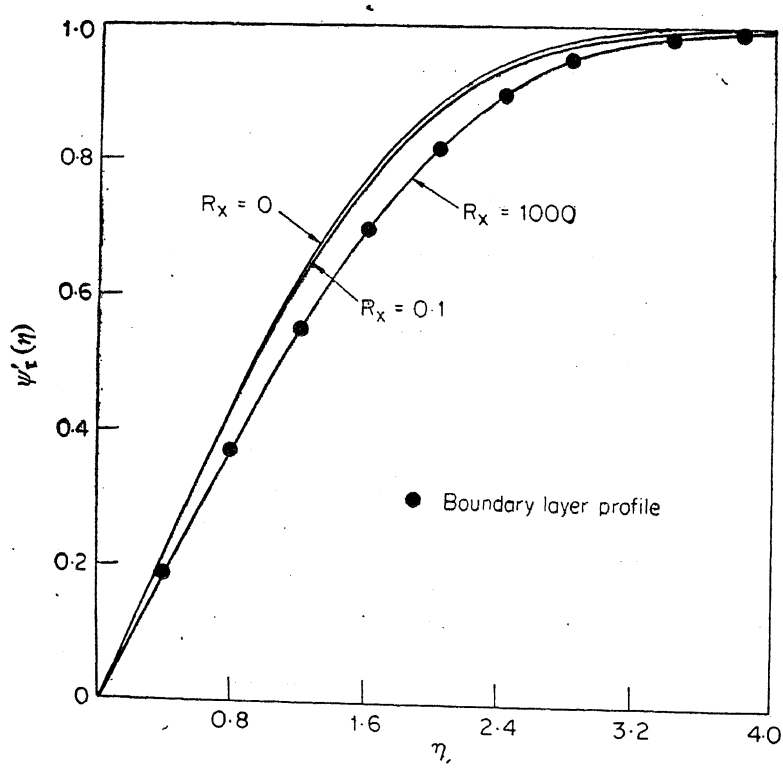

Figure 2. Velocity distribution at different values of R_x.

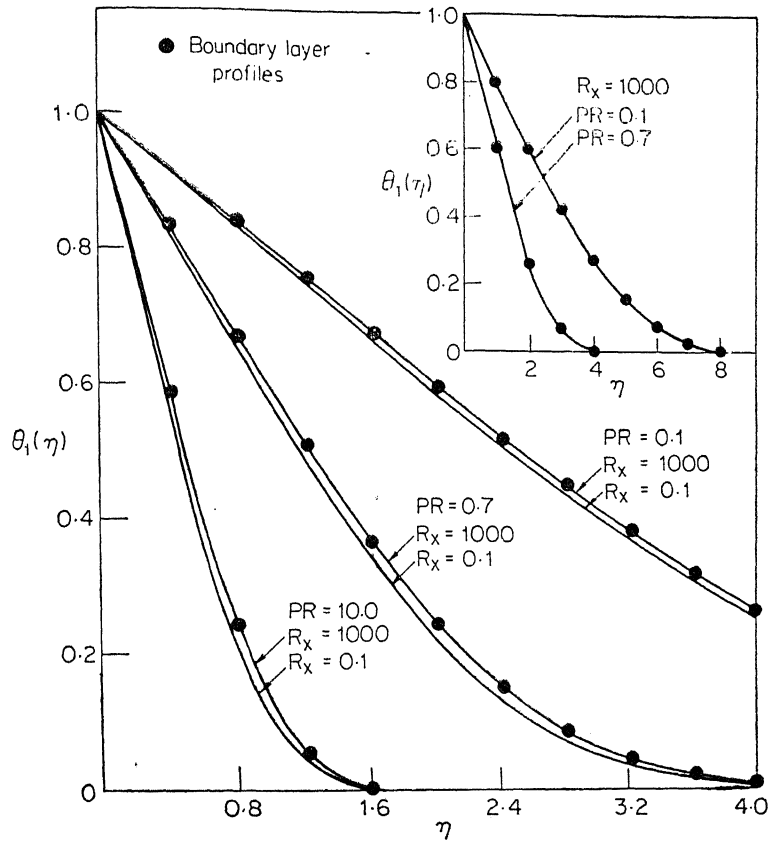

Figure 3. Temperature distribution at different values of R_x and Pr.

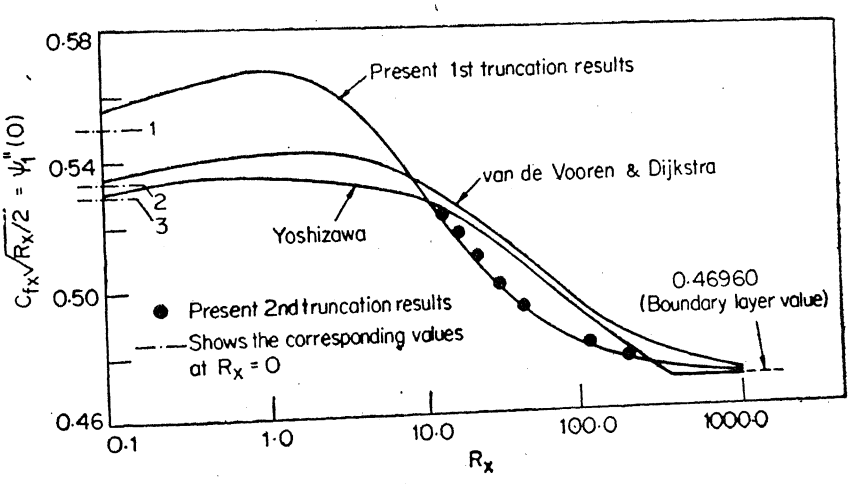

Figure 4. Effect of R_x on skin friction.

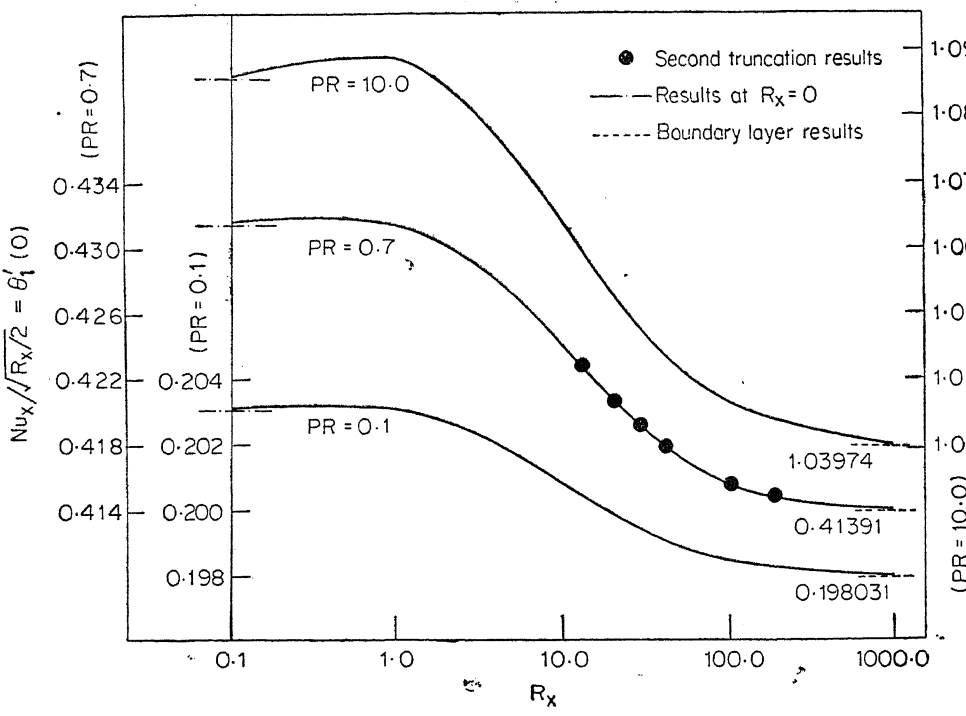

Figure 5. Effect of R_x on heat transfer.

linearly with $x^{1/2}$, however, for lower R_x it is a nonlinear function. The effect of the increase in Pr at a fixed R_x is to make the temperature profiles fuller and thus reduce δ_t.

Dimensionless local skin friction $[c_{fx}(R_x/2)^{1/2}]$ and dimensionless local heat transfer $[Nu_x/(R_x/2)^{1/2}]$ results are displayed in figures 4 and 5 respectively. The results indicate that though there is a little difference between first and second truncations for skin friction, there is much little difference for heat transfer. Skin friction results are same as those given by Davis [2] which compared favourably with those of Dean [4]. These results show that skin friction increases as R_x increases from its leading-edge value attaining a maximum around $R_x = 1{\cdot}2$; after which it decreases monotonically to boundary-layer value at about $R_x = 1000$. In figure 4 skin friction results have also been compared with the later studies of Van de Vooren and Dijkstra [9] and Yoshizawa [15] based on the solution of full partial Navier–Stokes equations. It may however be noted that there seems to be some discrepancy in van de Vooren and Dijkstra's paper in plotting Davis' results. It should also be noted that in Yoshizawa's work no variation in skin friction exists after $R_x = 338$, because his solution has been patched with the Blasius solution at this R_x; which must have some influence on these results. It is interesting to note that the leading-edge skin friction obtained by series truncation is only about 3% different than more exact results of van de Vooren and Dijkstra. This discrepancy is possibly because of exponential decay of stream function at large η, whereas the decay should be algebraic, see Davis [3].

The heat transfer results plotted in figure 5 show that boundary-layer values are attained around a R_x of 1000. For Pr = 0·1, and 0·7; heat transfer remains fairly constant upto a $R_x = 0{\cdot}4$, after which it decreases monotonically to corresponding boundary-layer values. However when Pr = 10, a distinct maximum heat transfer point around $R_x = 0{\cdot}6$ is observed in these results. The ratio between the values of

$$\underset{R_x \to 0}{\mathrm{Nu}_x/(R_x/2)^{1/2}} \quad \text{and} \quad \underset{R_x \to \infty}{\mathrm{Nu}_x/(R_x/2)^{1/2}}$$

is fairly ($\approx 1{\cdot}05$) constant for the various values of Pr investigated, and is smaller compared ($\approx 1{\cdot}17$) to the ratio between

$$\underset{R_x \to 0}{c_{fx}\,(R_x/2)^{1/2}} \quad \text{and} \quad \underset{R_x \to \infty}{c_{fx}\,(R_x/2)^{1/2}}.$$

This result implies that the effect of leading edge on heat transfer is smaller than in skin friction.

References

[1] Carrier G F and Lin C C 1948 *Q. Appl. Math.* **6** 63
[2] Davis R T 1967 *J. Fluid Mech.* **27** 691
[3] Davis R T 1972 *J. Fluid Mech.* **51** 417
[4] Dean W R 1954 *Mathematika* **1** 143
[5] Goldstein S 1960 *Lectures on fluid mechanics* (London : Interscience)
[6] Imai I 1957 *J. Aeronaut. Sci.* **24** 155
[7] Kaplun S 1954 *Z. Angew. Math. Phys.* **5** 111
[8] Murray J D 1965 *J. Fluid Mech.* **21** 337
[9] van de Vooren A I and Dijkstra D 1970 *J. Engg. Math.* **4** 9
[10] van Dyke M 1962 *J. Fluid Mech.* **14** 481
[11] van Dyke M 1964 *Perturbation methods in fluid mechanics* (New York : Academic Press)
[12] van Dyke M 1969 *Ann. Rev. Fluid Mech.* **1** 265
[13] van Dyke M 1967 *Fluid dynamics transactions* (ed.) W Fiszdon *et al* (Poland : Jurata) p. 53
[14] van Dyke M 1964 *Proc. 11th Int. Congr. Appl. Mech.* (Berlin : Springer Verlag)
[15] Yoshizawa A 1970 *J. Phys. Soc. Jpn.* **28** 776

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 2, May 1980, pp. 125–132.

On unsteady dispersion flow in porous media

MOHD. A A ANSARI
Department of Mathematics, Banaras Hindu University, Varanasi 221 005, India

MS received 22 August 1978; revised 16 April 1979

Abstract. The generalising dispersion equations of flow through porous media have been investigated. The Laplace transform has been applied to obtain the solution to dispersion problem as a result of adsorption. The generalised closed form solution for dispersion has been presented and the different types of variations in concentration have been graphically discussed. When the steady state occurs, the concentration becomes constant but for small value of time (say 0·5) the concentration tends to zero as distance increases.

Keywords. Adsorption; dispersion; porous media; seepage.

1. Introduction

Dispersion and simultaneous adsorption are common phenomena in the field of ion exchange in soils, salt water intrusion into coastal aquifers, underground disposal of biological, chemical and radiological agents, seepage from canals and streams into aquifers, and fluid removal in the petroleum reservoirs. Problems of dispersion and adsorption in porous media have attracted considerable attention from chemical engineers, environmental engineers, hydrologists, and petroleum engineers as well as soil scientists.

The present paper is concerned with the manner in which an originally sharp interface between two miscible fluids is altered as the fluids undergo unsteady motion in a porous medium. Banks and Jerasate [1] derived the governing equation for one-dimensional unsteady problem with adsorption within the aquifer by taking the relationship between dispersion coefficient that is nearly proportional to the product of the fluid velocity and particle diameter. This work was undertaken to study one aspect of the problem of salinity intrusion in ground water. Marino [4] also developed analytical solutions for two dispersion problems in non-adsorbing homogeneous porous media to study the possible contamination of ground water from seepage high salt concentration in drainage, ditches, canals and streams.

Bruce and Street [2] considered both longitudinal and lateral dispersion within a semi-infinite non-adsorbing porous medium in a unidirectional flow field. An analytical solution was developed for a constant input concentration. Ogata [6] presented various analytical solutions for dispersion problems within semi-infinite

adsorbing porous media in a unidirectional steady flow fields. The dispersion systems are subject to a constant input concentration.

In this paper, the dispersion in unsteady porous media flow has been considered and the mathematical solutions to two simplified dispersion problem involving retardation factor obtained by using the Laplace transform have been described. The solutions for concentration have been derived for $\omega = 0$ and $\omega \neq 0$. The concentration increases as time t increases and is constant in steady state. For the second case, the concentration decreases for small velocity and increases for large values of velocity.

2. Formulation of the problem

The theory is limited to unsteady longitudinal dispersion in one-dimensional seepage flows through semi-infinite, isotropic uniform porous media. The solutions developed are the concentration of the fluid phase as a function of time and space and solid phase as a function of time only. The concentration within the source is assumed to be varied with respect to time, although the seepage velocity distribution is one-dimensional. The concentration distribution is a result of longitudinal and adsorption within the aquifer, the equation for hydrodynamic dispersion within the semi-infinite medium in a unidirection flow field may be expressed as [1, 6]

$$\frac{\partial C}{\partial t} + \frac{1-n}{n}\frac{\partial F}{\partial t} + u\frac{\partial C}{\partial x} = D\frac{\partial^2 C}{\partial x^2}, \tag{1}$$

in which C is the concentration of the fluid phase, F is the concentration of the solid phase, D is the dispersion coefficient, u, the seepage velocity, n is porosity and t, the time.

Lapidus and Amundson [3] consider two cases namely,

$$F = K_1 C + K_2, \tag{2}$$

$$\partial F/\partial t = K_1 C - K_2 F, \tag{3}$$

representing, respectively, equilibrium and non-equilibrium, relationship between the concentration in the two phases. For simplicity, the former is adopted in the present analysis.

Introducing equation (2) into (1) and defining the adsorption coefficient R_d as

$$R_d = 1 + K_1\frac{1-n}{n}, \tag{4}$$

we have $$R_d\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} = D\frac{\partial^2 C}{\partial x^2}. \tag{5}$$

3. Method of solution

3.1. *Case I*

The system is schematically represented in figure 1. The concentration of the displacing fluid substance at $x = 0$ is $C_0 \exp(\omega t)$ in which ω is constant, positive or negative.

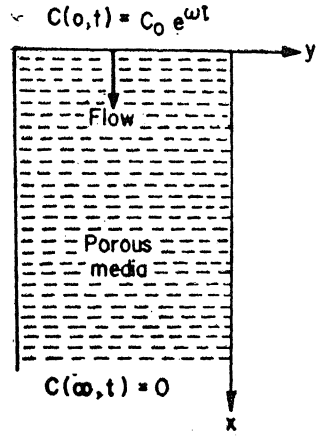

Figure 1. Semi-infinite porous medium in unidirectional flow field.

We can write the dispersion problem as equation (1) with initial and boundary conditions.

$$C(x, o) = 0, \qquad x \geqslant 0, \tag{6}$$

$$C(o, t) = C_0 e^{\omega t}, \quad t > 0, \tag{7}$$

$$C(\infty, t) = 0, \qquad t \geqslant 0. \tag{8}$$

For simplicity, we transform (5) so that the convective term does not appear, using the transformation [5]

$$C(x, t) = C_1(x, t) \exp\left(\frac{ux}{2D} - \frac{u^2 t}{4R_d D}\right), \tag{9}$$

and using equation (9), equations (5) to (8) can be written as,

$$R_d \frac{\partial C_1}{\partial t} = D \frac{\partial^2 C_1}{\partial x^2}, \tag{10}$$

$$C_1(x, O) = 0, \tag{11}$$

$$C_1(O, t) = C_0 \exp \eta t, \tag{12}$$

$$C_1(\infty, t) = 0, \tag{13}$$

where $$\eta = \frac{u^2}{4R_d D} + \omega. \tag{14}$$

Let $\bar{C}_1(x, S)$ be the transform with respect to t of the concentration function $C_1(x, t)$. Applying the Laplace transformation with respect to t on the preceding boundary value problem, we have

$$\frac{d^2 \bar{C}_1}{dx^2} - \frac{R_d}{D} S \bar{C}_1 = 0, \tag{15}$$

$$\bar{C}_1(O, S) = C_0/S - \eta, \tag{16}$$

$$\bar{C}_1 = (\infty, S) = 0 \tag{17}$$

in which the bars indicate the transformed function and S the parameter of the transformation.

The general solution of equation (15) is

$$\bar{C}_1(x, S) = A_1 \exp\left[-x(R_d S/D)^{1/2}\right] + A_2 \exp\left[x(R_d S/D)^{1/2}\right], \tag{18}$$

where A_1 and A_2 are constants of integration.

With the help of boundary conditions, equation (18) gives

$$A_1 = C_0/S - \eta, \tag{19}$$

$$A_2 = 0. \tag{20}$$

Therefore equation (18) becomes

$$\bar{C}_1(x, S) = \frac{C_0}{S-\eta} \exp\left[-x(R_d S/D)^{1/2}\right]. \tag{21}$$

Taking the inverse Laplace transformation equation (21) gives

$$C_1(x, t) = \tfrac{1}{2}C_0 \exp(\eta t)\left\{\exp\left[-x(R_d\eta/D)^{1/2}\right] \times \right.$$

$$\operatorname{erfc}\left[\frac{x\sqrt{R_d}}{2\sqrt{Dt}} - (\eta t)^{1/2}\right] +$$

$$\left.\exp\left[x(R_d\eta/D)^{1/2}\right] \operatorname{erfc}\left[\frac{x\sqrt{R_d}}{2\sqrt{Dt}} + (\eta t)^{1/2}\right]\right\}. \tag{22}$$

Finally, we get from equation (9)

$$C(x, t) = \tfrac{1}{2}C_0 \exp(\omega t)\left\{\exp\left[x(u-\xi)/2D\right] \times \right.$$

$$\operatorname{erfc}\left(\frac{xR_d - \xi t}{2(Dt R_d)^{1/2}}\right) + \exp\left[x(u+\xi)/2D\right] \times$$

$$\left.\operatorname{erfc}\left(\frac{xR_d + \xi t}{2(Dt R_d)^{1/2}}\right)\right\}, \tag{23}$$

where $$\xi = (u^2 + 4D R_1 \omega)^{1/2}. \tag{24}$$

Note that, if $\omega = 0$, equation (23) becomes

$$C(x, t) = \tfrac{1}{2}C_0\left\{\operatorname{erfc}\left(\frac{xR_d - ut}{2(Dt R_d)^{1/2}}\right) + \exp(ux/D) \times \right.$$

$$\left.\operatorname{erfc}\left(\frac{xR_d + ut}{2Dt R_d}\right)\right\}. \tag{25}$$

This result has been obtained by Marino [4] for non-adsorbing porous media, which has also been derived by Ogata and Banks [5] for the case in which the displacing fluid has a constant concentration C_0.

The second term of equation (25) is generally small. Thus, a good approximation to the solution may be written as

$$C(x, t) = \tfrac{1}{2}C_0 \operatorname{erfc}\left[\frac{xR_d - ut}{2(Dt R_d)^{1/2}}\right]. \tag{26}$$

3.2. *Case II*

The porous medium is represented in figure 1, except that the concentration of the displacing fluid at $x = 0$ is

$$C_0 (1 - e^{-\omega t}),$$

where C_0 is constant concentration at plate for $t \to \infty$.

Using the transformation indicated by equation (9), the hydrodynamic problem is reduced to:

$$D \frac{\partial^2 C_1}{\partial x^2} = R_d \frac{\partial C_1}{\partial t}. \tag{27}$$

The boundary conditions are reduced as

$$C_1 (x, 0) = 0, \tag{28}$$

$$C_1 (0, t) = C_0 [\exp\ (\alpha t) - \exp\ (\beta t)], \tag{29}$$

$$C_1 (\infty, t) = 0, \tag{30}$$

where $\alpha = u^2/4R_d D$, (31)

$$\beta = u^2/4R_d D^{-\omega}. \tag{32}$$

Now using the Laplace transformation in the usual manner to equations (27)–(30), we get

$$\frac{d^2 \bar{C}_1}{dx^2} - \frac{R_d}{D} S\bar{C}_1 = 0, \tag{33}$$

$$\bar{C}_1 (0, S) = C_0 \left[\frac{1}{S - \alpha} - \frac{1}{S - \beta}\right], \tag{34}$$

$$\bar{C}_1 (\infty, S) = 0. \tag{35}$$

Solving (33)–(35) we get

$$C(x, t) = \tfrac{1}{2} C_0 \left\{\operatorname{erfc}\left(\frac{xR_d - ut}{2(Dt R_d)^{1/2}}\right) + \exp\ (ux/D) \times \right.$$

$$\operatorname{erfc}\left(\frac{xR_d + ut}{2(Dt R_d)^{1/2}}\right) - \exp\ (-\omega t)\left[\exp\ \{x(u - \phi)/2D\} \times \right.$$

$$\operatorname{erfc}\left(\frac{xR_d - \phi t}{2(Dt R_d)^{1/2}}\right) + \exp\ (x(u + \phi)/2D) \times$$

$$\left.\left.\operatorname{erfc}\left(\frac{xR_d + \phi t}{2(Dt R_d)^{1/2}}\right)\right]\right\}, \tag{36}$$

where $\phi = (u^2 - 4R_d D\omega)^{1/2}$. (37)

This is the case when $\omega \neq 0$, where x is in the positive direction ($x > 0$). Note that for most practical purposes, the second and fourth terms of equation (36) can be neglected as [5]

$$C(x,t) = \tfrac{1}{2}C_0 \left\{ \operatorname{erfc}\left(\frac{xR_d - ut}{2(Dt\,R_d)^{1/2}}\right) - \exp(-\omega t) \times \right.$$

$$\left. \exp\frac{x(u-\phi)}{2D} \operatorname{erfc}\left(\frac{xR_d - \phi t}{2(Dt\,R_d)^{1/2}}\right)\right\}, \tag{38}$$

where $\phi = (u^2 - 4R_d D\omega)^{1/2}$, (39)

when $t \to \infty$ in equation (38), the concentration is constant in steady state.

4. Results

Solutions of equations (26) and (38) are presented graphically in figures 2 and 3 respectively for values of D and u representing those reported in [1] values of $D = 0{\cdot}008\ \text{cm}^2/\text{sec}$, $u = 0{\cdot}13\ \text{cm/sec}$ and $R_d = 1$. If we put $\omega = 0$ in equation (38) the concentration vanishes.

The values of concentration for

$$u = 0{\cdot}13\ \text{cm/sec},\ D = 0{\cdot}008\ \text{cm}^2/\text{sec},$$

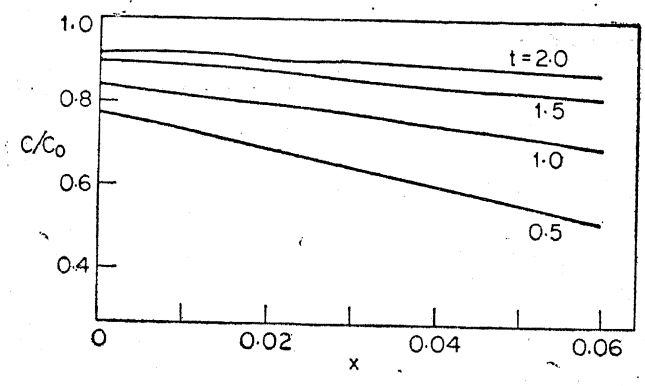

Figure 2. Concentration curves for numerical values of case 1.

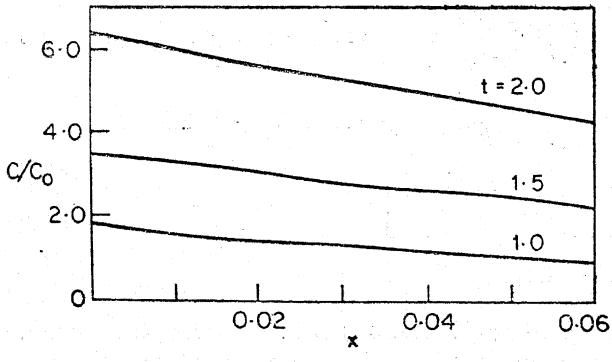

Figure 3. Concentration curves for numerical values of case 2.

Table 1. Values of C/C_0 for $\omega = 0$.

$x \backslash t$	0·5	1·0	1·5	2·0
0·00	0·7673	0·8413	0·9032	0·9251
0·01	0·7324	0·8212	0·8907	0·9177
0·02	0·6950	0·8023	0·8790	0·9082
0·03	0·6517	0·7794	0·8643	0·8979
0·04	0·6103	0·7549	0·8485	0·8888
0·05	0·5675	0·7324	0·8340	0·8790
0·06	0·5230	0·7054	0·8159	0·8665

Table 2. Values of C/C_0 for $\omega = 1$.

$x \backslash t$	0·5	1·0	1·5	2·0
0·00	0·1709	1·7533	3·5162	6·4101
0·01	0·1277	1·6068	3·2584	5·5981
0·02	0·997	1·4822	3·0656	5·6434
0·03	0·0659	1·3487	2·8307	5·2658
0·04	0·522	1·2451	2·6516	4·9627
0·05	0·0273	1·1267	2·4493	4·6197
0·06	0·0091	1·0180	2·2597	4·2980

and $R_d = 1$ are given above in tables 1 and 2 for equations (26) and (38) respectively.

Marino [4], has given equation (31) for the concentration as

$$C(x, t) = \tfrac{1}{2}C_0 \left\{ \operatorname{erfc}\left(\frac{x - ut}{2(Dt)^{1/2}}\right) - \exp(-\gamma t) \right.$$

$$\left. \exp\left[\frac{x(u - \phi)}{2Dt}\right] \operatorname{erfc}\left(\frac{x - \phi t}{2(Dt)^{1/2}}\right) \right\},$$

where ϕ is defined by equation (29) and other symbols have their usual meaning in the above reference.

If we use the values of ω as in table 2 in the present investigation, we find the concentration as negative.

Similar conclusion can be derived from table 2, when all values are negative. Positive values of ω cannot be taken since they are imaginary in equation (39), and therefore to avoid separation into real and imaginary part, ω has been taken as negative as was done by Marino [4].

5. Conclusions

Analytical solutions have been developed for two longitudinal dispersion problems in saturated porous media. The porous media considered is adsorbing, homogeneous and isotropic. The seepage flows are assumed to be unidirectional and the average velocities are taken to be constant throughout the flow fields.

From tables 1 and 2 it is clear that the concentration increases or decreases as t and x increases. The concentrations are symmetrical along t in increasing order and along x in decreasing order.

The concentration profiles are plotted in figures 2 and 3. In figure 2 for small t (say 0.5) the concentration decreases along x-axis, but for $t \geqslant 1{\cdot}0$ the concentration decreases slightly along the distance and the concentration increases as t increases and x decreases. From figure 3, the concentration distributions have the same behaviour as in figure 2 but for small t (say $0{\cdot}5$) the concentration tends to zero. When $\omega \neq 0$, equation (38) shows that the concentration decreases.

Acknowledgements

Thanks are due to Dr K Lal for his guidance. The present work was financially supported by Banaras Hindu University in the form of a fellowship.

References

[1] Banks R B and Jerasate S 1962 *J. Hydraul. Res.* **88** 1
[2] Bruce J C and Street R L 1966 Studies of free surface and two-dimensional dispersion in porous media. Rep. 63, Dept. of Civil Engg. Stanford Univ. Calif. p. 138
[3] Lapidus L and Amundson N R 1952 *J. Phys. Chem.* **56** 984
[4] Marino M A 1974 *J. Hydraul. Res.* **100** 151
[5] Ogata A and Banks R B 1961 A solution of the differential equation of longitudinal dispersion in porous media. Professional Paper No. 411–A–US Geol. Survey, Washington, DC
[6] Ogata A 1970 Theory of dispersion in a granular medium. US Geol. Surv. Professional Paper 411–I, p. 11
[7] Shen H T 1976 *J. Hydraul. Res.* **102** 707

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 3, September 1980, pp. 133–157.

Ramanujan and the congruence properties of partitions

K G RAMANATHAN
School of Mathematics, Tata Institute of Fundamental Research, Bombay 400 005, India

MS received 9 July 1980

Abstract. Ramanujan's unpublished manuscripts which came to light recently show that he had made significant advances towards proving the conjectures, on the congruence properties of the partition function, which were made by him. He also had proved several congruence relations for prime moduli other than 5, 7 and 11. In this paper a complete account of Ramanujan's work on the congruences for 5 and 7 and their powers and for the prime 13 is given.

Keywords. Congruence properties; partition function; Ramanujan's work.

1. Introduction

1.1. Let $p(n)$ denote the number of unrestricted partitions of the positive integer n so that

$$\sum_{n=0}^{\infty} p(n)\, x^n = \frac{1}{(1-x)(1-x^2)\ldots} = \Pi_n (1-x^n)^{-1}, \tag{1.11}$$

$|x| < 1$ and $p(0) = 1$. In 1918 Hardy and Ramanujan published their famous paper on the asymptotic formula for $p(n)$. In order to illustrate the agreement of this formula with results from actual computation they had a table of $p(n)$ from $n = 1$ to $n = 200$ prepared. Looking at the table of values of $p(n)$, Ramanujan wrote down several congruence properties of $p(n)$ for the moduli 5, 7 and 11 and their powers and then put forth the general conjecture to the effect that if a, b and c are non-negative integers and

$$m = 5^a \cdot 7^b \cdot 11^c, \qquad 24n \equiv 1 \pmod{m}, \tag{1.12}$$

then $$p(n) \equiv 0 \pmod{m}. \tag{1.13}$$

In particular, if $m = 5$, then $n = 4, 9, 14, \ldots$ so that

$$p(4),\ p(9),\ p(14),\ \ldots \equiv 0 \pmod 5.$$

In a beautiful note [10], Ramanujan gave proofs of the congruences for $m = 5$ and 7 and indicated a general method of proof by establishing identities of the type

$$\sum_{n=0}^{\infty} p(5n+4)\, x^n = 5\,\frac{\Pi(1-x^{5n})^5}{\Pi(1-x^n)^6}. \tag{1.14}$$

Although he indicated that he would come back to the general case of such identities, he unfortunately never did. In a still further note published in 1919 [11], at the time of his departure for India (27 February 1919), he mentions that he had proved (1.13) for $m = 11$ and that further there are congruences like

$$p(49n+19),\ p(49n+33),\ p(49n+40),\ p(49n+47) \equiv 0 \pmod{49}.$$

Note that except for the last one, the other three do *not* come under (1.12).

Sometime after his election to Fellowship of Trinity College, Cambridge (13 October 1918), Ramanujan wrote to Hardy (see Appendix) that he had 4 methods for studying the congruence properties of $p(n)$ and the more general function $p_{-k}(n)$, the method by theta functions being the one indicated in [10]. Later Ramanujan wrote a long paper [14], which was never published, consisting of 43 foolscap size pages and entitled "*Properties of $p(n)$ and $\tau(n)$ defined by the equations*

$$\sum_{n=0}^{\infty} p(n)\, x^n = \frac{1}{(1-x)(1-x^2)(1-x^3)\cdots}$$

$$\sum_{n=1}^{\infty} \tau(n)\, x^n = x\{(1-x)(1-x^2)(1-x^3)\ldots\}^{24}.$$

In it he had developed his methods of Dirichlet series and the functions Q and R and had given proofs of (1.13) for $a = 0, 1, 2, 3$; $b = 0, 1, 2$; and $c = 0, 1, 2$. He says, in his letter to Hardy, that he had proved (1.13) for $b = 3$. This is, however, not true. It is known [22] that for $m = 7^3$ Ramanujan's conjecture is false. The manuscript [14] was in the nature of a first draft of a long paper on congruence properties of certain arithmetical functions. Hardy received this manuscript in 1920 perhaps a few months before Ramanujan's death as Rushforth says or after Ramanujan's death as B M Wilson says, along with other manuscripts. In any case Hardy, according to G K Stanley, intended to edit it [20]. He however extracted and posthumously published [12] Ramanujan's proof of (1.13) for $m = 11$. The proof made use of Ramanujan's functions Q and R. The manuscript was never edited and seems to have remained with G N Watson. J M Rushforth wrote with Watson a thesis [19] in which he proved Ramanujan's statements on $p(n)$ related to the moduli 11^2, 13, 17, 19, 23, 29 and 31. The proofs again used Ramanujan's functions Q and R.

Ramanujan had another manuscript [15] which is not available now, parts of which had however been copied down by G N Watson (see [5]) and are now in the Oxford Mathematical Library. The originals, perhaps, have perished. In it Ramanujan ([15], pp. 252–257) gives a fairly complete proof of (1.13) for $m = 5^a$ for all $a \geqslant 1$ and for $m = 7^\beta$, $\beta = 1, 2$. The proof is entirely different from that in [14]. This manuscript of Ramanujan's seems to be the part II of the

paper [14]. In [14], p. 27 Ramanujan says " ... but I shall consider in the II par of this paper the analogous formulae for the smaller primes 5, 7 and 11 ". The paper [14] contains 19 sections and [15] p. 252 begins with § 20 with the words " In this second part we shall use ... ". Further the manuscript [14] has 34 pages with 9 pages of insertions and Watson's note [15] rightly begins with page 35. Watson's manuscript [15] is thus part II of [14] and is again incomplete. Ramanujan's and Watson's [22] proofs of (1.13) for $m = 5^{\alpha}$ or 7^{β} depend on the modular equations of degrees 5 and 7. These modular equations are stated in the Note books ([13], p. 235 and p. 239) in the irrational form. A proof, in the case of 5, can be extracted from [13] and [15].

Except for the proofs of (1.13) for the moduli 11 and 11^2 contained in [14], there is no indication, in any of the unpublished manuscripts, of any general procedure for the modulus 11^{γ}, $\gamma \geqslant 3$. Even in his letter to Hardy, he stops short of 11^2. As is well-known, the complete proof for 11^{γ}, $\gamma \geqslant 1$ was given by A O L Atkin [3].

As for congruences for higher prime moduli, Ramanujan proved in [14] the congruences

$$\sum_{n=0}^{\infty} p(13n + 6)\, x^n \equiv 11 \prod (1 - x^n)^{11} \pmod{13}, \tag{1.15}$$

$$\sum_{n=0}^{\infty} p(13^2 n + 162)\, x^n \equiv 10 \prod (1 - x^n)^{23} \pmod{13}, \tag{1.16}$$

by using the functions Q and R. He also remarks " since I began to write this paper I have found by a different method, that if λ be any positive odd integer then

$$\sum_{n=0}^{\infty} p\left(13^{\lambda} n + \frac{11.13^{\lambda} + 1}{24}\right) x^n = -2^{(5\lambda-3)/2} \{(1 - x)(1 - x^2) \ldots\}^{11} + 13J \tag{1.17}$$

and if λ is any positive even integer then

$$\sum_{n=0}^{\infty} p\left(13^{\lambda} n + \frac{23.13^{\lambda} + 1}{24}\right) x^n = -2^{(5\lambda-2)/2} \{(1 - x)(1 - x^2) \ldots\}^{23} + 13J, \tag{1.18}$$

where J is a power series in x with integer coefficients ". He says on page 21 of [14] " I shall reserve the discussion of these results to another paper ". Unfortunately he passed away before he could fulfil his promise.

Recently some new and unpublished manuscripts of Ramanujan's have come to light [16]. They are called by G E Andrews [1], who first drew my attention to them, the " Lost " note book. In this manuscript, whose pagination is rather haphazard, there are just 3 pages which contain some results relating to congruences of $p(n)$ and $p_r(n)$. This manuscript was written towards the end of Ramanujan's life in India. It shows what an enormous amount of work he had done during the last year of his life in spite of his illness.

The object of our paper is to present Ramanujan's work on $p(n)$ and $p_r(n)$ and to give, in a simplified form Ramanujan's methods and results in [15] for

the modulus 5^α and a similar proof for 7^β. Our method is, to a certain extent, inspired by Atkin. Later we make some remarks regarding Ramanujan's results (1.15)–(1.17) in the light of results in the "Lost" note book [16]. We also point out their relation to the work of Morris Newman [7] and [8], who independently proved many results of Ramanujan's in [14]. Finally we make some remarks on the three pages in [16] mentioned above.

We shall report on another occasion on Ramanujan's results on congruences to prime moduli greater than 13 contained in the MS [14]. This manuscript also contains a large number of results on the function $\tau(n)$ which need to be reported on especially in view of the recent work of Serre and Swinnerton–Dyer.

2. Modulus 5^a

2.1. Let $|x| < 1$ and

$$f(x) = \prod_{n=1}^{\infty} (1 - x^n). \tag{2.11}$$

Then $$f(x) = 1 + \sum_{n=1}^{\infty} (-1)^n (x^{n(3n-1)/2} + x^{n(3n+1)/2}), \tag{2.12}$$

and $$(f(x))^3 = \sum_{n=0}^{\infty} (-1)^n (2n+1)\, x^{n(n+1)/2}. \tag{2.13}$$

The first identity is due to Euler and the second to Jacobi [6].

Let us put

$$\varphi_5(x) = \varphi(x) = x \frac{f(x^{25})}{f(x)}$$

$$g_5(x) = g(x) = x \frac{f^6(x^5)}{f^6(x)}. \tag{2.14}$$

The modular equation of degree 5 is the polynomial relation between $\varphi(x^{1/5})$ and $g(x)$. Ramanujan states this relation in his Note books [[13] Vol. II, p. 235] without proof. It is in the following form : If $2u = 11 + 1/g(x)$ and $2v = 1 + 1/\varphi(x^{1/5})$, then

$$[(u^2+1)^{1/2} - u]^{1/5} = (v^2+1)^{1/2} - v.$$

"Rationalising" the surds we get

$$t^5 = g(x)(25t^4 + 25t^3 + 15t^2 + 5t + 1) \tag{2.15}$$

where $t = \varphi(x^{1/5})$. This is the modular equation of degree 5 whose roots, as a polynomial in t are $\varphi(\rho x^{1/5})$ where ρ runs through the different 5th roots of unity. A proof of (2.15) can be extracted from the manuscript [15]. From (2.12) we see that

$$f(x^{1/5})/f(x^5) = J_1 - x^{1/5} - x^{2/5} J_2,$$

where J_1 and J_2 are power series in x with integer coefficients. If we cube both sides and use (2.13), we get

$$\frac{\sum_{n=0}^{\infty}(-1)^n(2n+1)x^{n(n+1)/10}}{[f(x^5)]^3} = (J_1 - x^{1/5} - x^{2/5}J_2)^3.$$

Since for any n, $n(n+1) \equiv 0, 2$ or $6 \pmod{10}$, there is no term on the left containing $x^{2/5}$. Therefore the term with $x^{2/5}$ on the right (when expanded) is zero. Thus

$$3J_1(1 - J_1J_2) = 0,$$

or
$$J_1J_2 = 1. \tag{2.16}$$

This gives

$$[\varphi(x^{1/5})]^{-1} = (J_1/x^{1/5}) - 1 - (x^{1/5}/J_1). \tag{2.17}$$

Putting $\rho x^{1/5}$ instead of $x^{1/5}$ in (2.17) and multiplying for all fifth roots of unity ρ, we get

$$\Pi_\rho\,[\varphi(\rho x^{1/5})]^{-1} = \Pi_\rho\,[(J_1/\rho x^{1/5}) - 1 - (\rho x^{1/5}/J_1)].$$

Using (2.11) we get

$$[g(x)]^{-1} = (J_1/x^{1/5}) - 11 - (x^{1/5}/J_1). \tag{2.18}$$

Ramanujan's equation is obtained by "solving" for $J_1/x^{1/5}$ from (2.17) and (2.18). However (2.15) is a simple consequence of these two results. For if $y = J_1/x^{1/5}$, then

$$11 + [g(x)]^{-1} = y^5 - \frac{1}{y^5} = \left(y - \frac{1}{y}\right)\left(y^4 + \frac{1}{y^4} + y^2 + \frac{1}{y^2} + 1\right).$$

But

$$y^2 + \frac{1}{y^2} = \left(y - \frac{1}{y}\right)^2 + 2,$$

$$y^4 + \frac{1}{y^4} = \left(y^2 + \frac{1}{y^2}\right)^2 - 2.$$

Using the fact that

$$y - \frac{1}{y} = 1 + [\varphi(x^{1/5})]^{-1},$$

we get (2.15).

2.2. We define the operator $U_5 = U$ which acts on power series $A(x)$ in x such that

$$U_5(A(x)) = U[A(x)] = \frac{1}{5}\sum_\rho A(\rho x^{1/5}),$$

where ρ runs over all the fifth roots of unity. Clearly for any two power series $A(x)$ and $B(x)$

$$U(A(x) + B(x)) = U(A(x)) + U(B(x)),$$

$$U(A(x) \cdot B(x^5)) = B(x) \cdot UA(x).$$

Further if $A(x)$ is a power series with integer coefficients, then so is $UA(x)$.

We shall give a complete proof of Ramanujan's congruences for $p(n)$ modulo 5^a for $a \geqslant 1$. This is contained in [15]. Indeed the inductive method which Watson uses in [22] is already contained in [15]. This is also the basis for the congruences mod 7^β.

Before we give the main results of Ramanujan's in [15] we give a rather simple proof of

$$p(5n + 4) \equiv 0 \pmod 5,$$

given by Ramanujan in [14]. Ramanujan states the identity

$$\psi(x) = \frac{x[f(x^5)]^5}{f(x)}$$

$$= \sum_{n=0}^{\infty} \left\{ \frac{x^{5n+1}}{(1-x^{5n+1})^2} + \frac{x^{5n+4}}{(1-x^{5n+4})^2} - \frac{x^{n5+2}}{(1-x^{5n+2})^2} - \frac{x^{5n+3}}{(1-x^{5n+3})^2} \right\}. \tag{2.21}$$

We apply the operator U_5 to both sides of (2.21). The left hand side is

$$U\psi(x) = [f(x)]^5 \sum_0^\infty p(5n + 4)\, x^{n+1}.$$

On the right hand side the operator can be applied to each term and we get if 5 does not divide n.

$$U\left(\frac{x^n}{(1-x^n)^2}\right) = U\left(\sum_{m=1}^{\infty} m x^{mn}\right) = 5\frac{x^n}{(1-x^n)^2}.$$

We therefore have

$$[f(x)]^5 \sum_0^\infty p(5n + 4)\, x^{n+1} = 5x\frac{[f(x^5)]^5}{f(x)}. \tag{2.22}$$

which is one of the beautiful identities of Ramanujan's.

2.3. We now commence the proof of the congruences. The idea, as Ramanujan indicated in [10] is to prove identities of type (2.22) which would make the congruences evident.

Since $\varphi(\rho x^{1/5})$ are roots of the modular equation (2.15),

$$U\varphi(x) = \frac{1}{5}\sum_\rho \varphi(\rho x^{1/5}) = \frac{1}{5} \cdot 25 \cdot g = 5g.$$

From the definition of $\varphi(x)$, we get

$$x \prod (1-x^{5n}) \sum_0^\infty p(5n + 4)\, x^n = 5x\frac{f^6(x^5)}{f^6(x)}, \tag{2.31}$$

which gives the Ramanujan identity

$$\sum_0^\infty p(5n+4)\,x^n = 5\frac{(1-x^{5n})^5}{(1-x^n)^6}. \tag{2.32}$$

This proves the Ramanujan conjecture for $a = 1$.

Ramanujan now says " changing x into $x^{1/5}$ in (2.31) and arguing as before... " ([15], p. 253). What Ramanujan means is that we should apply the operator U to both sides of (2.31), as he does in [15]. If we apply U to the left side of (2.31) we get

$$x\Pi(1-x^n)\sum_0^\infty p(5^2n+5.4+4)\,x^n.$$

Ramanujan thus writes down

$$x\Pi(1-x^n)\sum_0^\infty p(25n+24)\,x^n = 5^2\cdot 63.g + 5^5\cdot 52\cdot g^2 + 5^7\cdot 63.g^3 + 5^{10}\cdot 6\cdot g^4 + 5^{12}\cdot g^5, \tag{2.33}$$

which, of course, proves that $p(25n+24) \equiv 0 \pmod{25}$. This formula is explicitly written down by Ramanujan ([15], p. 253). It is also curious to note that on top of one of the three pages in [16], Ramanujan just writes down the numbers

$$5^2\cdot 63,\quad 5^5\cdot 52,\quad 5^7\cdot 63,\quad 5^{10}\cdot 6,\quad 5^{12}$$

without comment. These are, of course, the coefficients in (2.33).

After deriving (2.33), Ramanujan again says " changing x into $x^{1/5}$ ... ". The induction process is now obvious; if we observe that on the right of (2.31) we have $5g(x)$ and we have to apply U on $g(x)$ and verify (this Ramanujan asserts) that $Ug(x)$ is a polynomial in $g(x)$ with integer coefficients which are divisible by appropriate powers of 5. At the next stage one has to multiply both sides of (2.33) by $\varphi(x)$ and then apply the operator U. In order to continue the inductive step, it is necessary to study

$$U((g(x))^k), \qquad U(\varphi(x)\,g(x)^k),$$

for $k \geqslant 1$ and show that they are polynomials in $g(x)$ with integer coefficients.

Let us observe that

$$U(g(x)^k) = \frac{1}{5}\sum_\rho\left(\rho^k\cdot x^{k/5}\frac{f^{6k}(x)}{f^{6k}(\rho x^{1/5})}\right) = \frac{1}{5}\left(\frac{f^6(x)}{x f^6(x^5)}\right)^k\sum_\rho\left(\frac{\rho x^{1/5} f(x^5)}{f(\rho x^{1/5})}\right)^{6k}$$

$$= g^{-k}\,U(\varphi(x))^{6k}. \tag{2.34}$$

In a similar manner, we get

$$U[\varphi(x)\,g(x)^k] = g^{-k}\,U\{[\varphi(x)]^{6k+1}\}. \tag{2.35}$$

It is therefore enough to study $U[\varphi(x)]^k$ for any integer $k \geqslant 1$. Notice that $U[\varphi(x)]^k$ is upto the multiple 1/5, the sum of the kth powers of the roots of (2.15)

and from the form of (2.15) it will be a polynomial in g. We therefore have to use Newton's rule for the sums of powers of roots of a polynomial. We thus have the

Lemma 1. *For every integer* $k \geqslant 1$,

$$U_5[\varphi(x)]^k = \sum_{t \geqslant l} a_{kt} 5^{t-l} [g(x)]^t,$$

where a_{kt} *are (positive) integers vanishing for all large* t *and* $l = [(k+4)/5]$.

Proof: For $k \leqslant 5$, it can be verified directly from the modular equation using the formula for the sums of powers of roots of a polynomial.

Let $k > 5$ and assume that the lemma is proved for $k < m$. Put

$$U_5[\varphi(x)]^k = \sum_{t \geqslant l = \left[\frac{k+4}{5}\right]} lo_{kt}\, g(x)^t \tag{2.36}$$

for $k < m$ and

$$v(b_{kt}) \geqslant t - l, \quad k < m, \quad t \geqslant [(k+4)/5],$$

where, for any rational number a, $v(a)$ is the highest power of 5 dividing a. Let now $k = m$. From (2.15) we have

$$\begin{aligned} U(\varphi(x))^m &= 25gU(\varphi(x))^{m-1} + 25gU(\varphi(x))^{m-2} + 15gU(\varphi(x))^{m-3} \\ &\quad + 5gU(\varphi(x))^{m-4} + gU(\varphi(x))^{m-5} \end{aligned} \tag{2.37}$$

where we write g instead of $g(x)$. Substituting, $1 \leqslant i \leqslant 5$

$$U[\varphi(x)]^{m-i} = \sum_{t \geqslant [(m-i+4)/5]} b_{m-i,t} [g(x)]^t$$

where $b_{m-i,t}$ are integers. Further, for all i

$$t \geqslant \min_{1 \leqslant i \leqslant 5} [(m-i+4)/5] \geqslant [(m+4)/5] - 1.$$

Therefore (2.36) and (2.37) show that $U(\varphi(x)^m)$ has the form (2.36) with b_{kt} integers and

$$t \geqslant [(m+4)/5].$$

Furthermore, for $t \geqslant [(m+4)/5]$

$$b_{mt} = 25b_{m-1,t-1} + 25b_{m-2,t-1} + 15b_{m-3,t-1} + 5b_{m-4,t-1} + b_{m-5,t-1},$$

so that $v(b_{mt}) \geqslant \min\{2 + v(b_{m-1,t-1}),\ 2 + v(b_{m-2,t-1}),\ 1 + v(b_{m-3,t-1}),$

$$1 + v(b_{m-4,t-1}),\ v(b_{m-5,t-1})\}.$$

By induction hypothesis, therefore

$$v(b_{mt}) \geqslant t - [(m+4)/5]$$

which proves the lemma.

In view of the next lemma, let us write down $U[g(x)]$ and $U[\varphi(x) \cdot g(x)]$. Since

$$U[g(x)] = g^{-1} U[\varphi(x)]^6,$$

we see, from direct computation

$$U[g(x)] = g^{-1} \sum_{t \geqslant 2} a_{6,t} 5^{t-2} g^t,$$

with $a_{6,t} \equiv 0 \pmod 5$, $t \geqslant 2$. (2.38)

In a similar way

$$U[\varphi(x) g(x)] = g^{-1} U[\varphi(x)]^7$$
$$= g^{-1} \sum_{t \geqslant 2} a_{7,t} 5^{t-2} g^t$$

and $a_{7,t} \equiv 0 \pmod 5$, $t \geqslant 2$. (2.39)

2.4. Let us define, inductively

$$L_1(x) = U_5[\varphi(x)],$$

and for $n > 1$

$$L_n(x) = U\{[\varphi(x)]^\epsilon L_{n-1}(x)\} \tag{2.41}$$

where $\epsilon = 0$ if n is even and $\epsilon = 1$ if n is odd. Also put $L_0(x) = 1$. We then have

Lemma 2. *For all* $k \geqslant 1$

$$L_k(x) = \sum_{t \geqslant 1} c_{k,t} 5^{t-1} g^t,$$

where $c_{k,t}$ *are integers vanishing for all large* t.

Proof: We prove this by induction on k. In case $k = 1$, $L_1(x) = U[\varphi(x)]$ and lemma 1 gives the result. Let us assume that lemma 2 is proved for all $k < m$. Let m be even; then

$$L_m(x) = U[L_{m-1}(x)]$$
$$= \sum_{p \geqslant 1} c_{m-1,p} 5^{p-1} U[g(x)]^p.$$

Using lemma 1, we get

$$L_m(x) = \sum_{p \geqslant 1} c_{m-1,p} 5^{p-1} \Big(\sum_{s \geqslant r} a_{6p,s} 5^{s-r} g^s\Big) g^{-p}, \tag{2.42}$$

where $r = [(6p + 4)/5] > p$.

We rearrange the summation and write (2.42) as

$$\sum_{t \geqslant 1} \Big(\sum_{s-p=t} c_{m-1,p} a_{6p,s} 5^{s-r+p-1}\Big) g^t \tag{2.43}$$

where, in the inner sum $p \geqslant 1$ and $s \geqslant [(6p + 4)/5] = r$. The inner sum is finite since $c_{m-1,p}$ and $a_{6p,s}$ are zero for all sufficiently large p and s and p is clearly bounded. Since $s \geqslant [(6p + 4)/5] = r$, and

$$2p \geqslant p + [(p + 4)/5] = [(6p + 4)/5] = r,$$

and in general for $k \geqslant 1$

$$\mu_{2k} = 4 + 4\cdot 5 + 3\cdot 5^2 + \ldots + 4\cdot 5^{2k-1}$$

$$\mu_{2k+1} = 4 + 4\cdot 5 + 3\cdot 5^2 + \ldots + 3\cdot 5^{2k}$$

which gives

$$\mu_{2k} = (23\cdot 5^{2k} + 1)/24$$

$$\mu_{2k+1} = (19.5^{2k+1} + 1)/24 .$$

In the corollary we see that the cases $\lambda = 2$ and 4 do not come under the theorem 1.

Ramanujan ([15], pp. 255–256) determined the residues of $p(5^k n + \mu_k) \bmod 5^{k+1}$. He finds the residues for $n = 1, 2, \ldots, 16$ for k odd and for $n = 1, \ldots, 16$ for k even. We shall prove his general statement.

We shall determine the residue of $c_{m,1}$ modulo 5^{m+1}. From (2.44) and (2.45) we have

$$c_{m,2} = \begin{cases} \sum_{s-p=2} c_{m-1,p}\, a_{6p,s}\, 5^{2p-r}, & m \text{ even} \\ \sum_{s-p=2} c_{m-1,p}\, a_{6p+1,s}\, 5^{2p-r}, & m \text{ odd} \end{cases}$$

where $s \geqslant [(6p+4)/5] = r$ in the first case and $s \geqslant [(6p+5)/5] = r$ in the second case. Since $s \geqslant 3$ in both cases, we find by actual computation

$$a_{6,3} \equiv 0 \pmod{5^2}$$

$$a_{7,3} \equiv 0 \pmod{5^2}. \tag{2.56}$$

We now prove

$$c_{m,2} \equiv 0 \pmod{5^{m+1}}. \tag{2.57}$$

Let first m be even. Then

$$c_{m,2} \equiv c_{m-1,1}\, a_{6,3} + c_{m-1,2}\, a_{12,4}\, 5 \pmod{5^{m+1}}.$$

Since $c_{m-1,1} \equiv c_{m-1,2} \equiv 0 \pmod{5^{m-1}}$, using (2.56) and the fact that $a_{12,4} \equiv 0 \pmod 5$ we obtain (2.57) for m even.

If m is odd, then

$$c_{m,2} \equiv c_{m-1,1}\, a_{7,3} + c_{m-1,2}\, a_{13,4} \cdot 5 \pmod{5^{m+1}}.$$

Since m is odd, $m-1$ is even and (2.56) and (2.57) give again (2.57) for m odd

Let us now write

$$c_{m,1} = \begin{cases} \sum_{s-p=1} c_{m-1,p}\, a_{6p,s}\, 5^{2p-r}, & m \text{ even} \\ \sum_{s-p=1} c_{m-1,p}\, a_{6p+1,s}\, 5^{2p-r}, & m \text{ odd} \end{cases}$$

By (2.57) we see that mod 5^{m+1}

$$c_{m,1} \equiv \begin{cases} c_{m-1,1}\, a_{6,2}, & m \text{ even} \\ c_{m-1,1}\, a_{7,2}, & m \text{ odd,} \end{cases}$$

By computation we find

$$a_{6,2} \equiv a_{7,2} \equiv 15 \pmod{5^2}.$$

We therefore obtain

$$a_m \equiv 3a_{m-1} \pmod 5. \tag{2.58}$$

Since $a_1 = 1$, we get

Corollary 2. *For all k, the congruences*

$$\sum_{n=0}^{\infty} p(5^k n + \mu_k) x^n \equiv 3^{k-1} \cdot 5^k \prod (1 - x^n)^{19}$$

for k odd and for k even

$$\sum_{n=0}^{\infty} p(5^k n + \mu_k) x^n \equiv 3^{k-1} \cdot 5^k \prod (1 - x^n)^{23}$$

modulo 5^{k+1} *hold.*

Ramanujan expands $\prod (1 - x^n)^{19}$ as apower series in x and thus gives the residues. If we use the notation

$$\prod (1 - x^n)^r = \sum_{n=0}^{\infty} p_r(n) x^n \tag{2.59}$$

where r is an integer $\neq 0$, then we may write the above corollary as

Corollary 2′

$$p(5^k n + \mu_k) \equiv \begin{cases} 3^{k-1} \cdot 5^k \cdot p_{19}(n), & k \text{ odd} \\ 3^{k-1} \cdot 5^k \cdot p_{23}(n), & k \text{ even} \end{cases}$$

the congruences being taken mod 5^{k+1}.

3. Modulus 7^b

3.1. Ramanujan's results on the congruences modulo 7^b are incomplete. He proves in [14], by using the familiar functions, P, Q, R, the congruence properties of $p(n)$ for the moduli 7 and 7^2. In [15] he again gives a nice proof of the identity

$$p(5) + p(12)x + p(19)x^2 + \ldots =$$

$$= 7\frac{\prod (1 - x^{7n})^3}{\prod (1 - x^n)^4} + 49x\frac{\prod (1 - x^{7n})^7}{\prod (1 - x^n)^8}, \tag{3.11}$$

by using the identity

$$x\prod (1 - x^n)^3 \prod (1 - x^{7n})^3 + 8x^2\frac{\prod (1 - x^{7n})^7}{\prod (1 - x^n)}$$

$$= \sum_{n=0}^{\infty} \left\{ \frac{x^{7n+1}(1 + x^{7n+1})}{(1 - x^{7n+1})^2} + \frac{x^{7n+2}(1 + x^{7n+2})}{(1 - x^{7n+2})^2} \right.$$

$$+\frac{x^{7n+4}(1+x^{7n+4})}{(1-x^{7n+4})^2}-\frac{x^{7n+3}(1+x^{7n+3})}{(1-x^{7n+3})^2}-\frac{x^{7n+5}(1+x^{7n+5})}{(1-x^{7n+5})^2}$$

$$-\frac{x^{7n+6}(1+x^{7n+6})}{(1-x^{7n+6})^2}\Big\}. \tag{3.12}$$

As before we define the operator $U_7 = U$ acting on power series in x such that

$$U_7 A(x) = \frac{1}{7}\sum_{\rho} A(\rho x^{1/7}),$$

ρ running through all the 7th roots of unity. U_7 has properties similar to U_5. Ramanujan's identity (3.11) is obtained by applying U_7 to both sides of (3.12).

Ramanujan gives, in his elementary way, all the preliminaries for writing down the modular equation of degree 7 which is the polynomial relation connecting $\varphi(x^{1/7})$ and $g(x)$ where

$$g(x) = g_7(x) = x f^4(x^7)/f^4(x),$$

$$\varphi(x) = \varphi_7(x) = x^2 f(x^{49})/f(x).$$

Ramanujan gives this relation explicitly, in his Note books ([13] Vol. II, p. 239) in the form

$$2u = 7(v^3 + 5v^2 + 7v) + (v^2 + 7v + 7)(4v^3 + 21v^2 + 28v)^{1/2}$$

where $u = [g(x)]^{-1}$, $v = [\varphi(x^{1/7})]^{-1}$.

After clearing radicals, one obtains the modular equation

$$t^7 = (7^2 g + 7^3 g^2)t^6 + (5\cdot 7 g + 7^3 g^2)t^5 + (7g + 3\cdot 7^2 g^2)t^4$$

$$+ 7^2 \cdot g^2 t^3 + 7^3 \cdot g^2 t^2 + 7 \cdot g^2 \cdot t + g^2. \tag{3.13}$$

The roots of this polynomial in t are $\varphi(\rho x^{1/7})$ where ρ runs through all the 7th roots of unity. Watson [22] proves (3.13) following the sketch given by Ramanujan in [15], pp. 256–257.

3.2. As in the case of the modulus 5, $\varphi(\rho x^{1/7})$ are the roots of the modular equation (3.13) and

$$U_7[\varphi(x)] = U[\varphi(x)] = \frac{1}{7}\cdot[7^2 g(x) + 7^3 g^2(x)],$$

from (3.13). This gives Ramanujan's identity

$$\Pi(1-x^{7n})\sum_0^{\infty} p(7n+2)x^n = 7\cdot\frac{\Pi(1-x^{7n})^4}{\Pi(1-x^n)^4}$$

$$+ 49x\frac{\Pi(1-x^{7n})^8}{\Pi(1-x^n)^8}. \tag{3.21}$$

Following Ramanujan's idea for the modulus 5^a we have the following lemmas.

Lemma 4. *For any integer* $k \geqslant 1$,

$$U_7[\varphi(x)]^k = \sum_{t\geqslant r} a_{k,t}\, 7^{t-r}[g(x)]^t,$$

where $r = [(2k+6)/7]$ *and* $a_{k,t}$ *are integers vanishing for all large* t.

By actual computation using the modular equation, we get

$$U[\varphi(x)]^4 = \sum_{t\geqslant 2} a_{4,t}\, 7^{t-2}\,[g(x)]^t,$$

where $a_{4,2} = 82.7 \equiv 0 \pmod 7$. (3.22)

and $$U[\varphi(x)]^5 = \sum_{t\geqslant 2} a_{5,t}\, 7^{t-2}\,[g(x)]^t,$$

with $a_{5,2} = 190 \not\equiv 0 \pmod 7$. (3.23)

As before we define, inductively

$$L_1(x) = U_7[\varphi(x)],$$

and for any $n > 1$

$$L_n(x) = U_7\{[\varphi(x)]^\varepsilon\, L_{n-1}(x)\},$$

where $\varepsilon = 1$ if n is odd and $= 0$ if n is even. We then have, as before,

Lemma 5. *For all integers* $k \geqslant 1$

$$L_k(x) = \sum_{t\geqslant 1} c_{k,t}\, 7^{t-1}\,[g(x)]^t,$$

$c_{k,t}$ *being integers (positive) vanishing for all sufficiently large* t *(for each given* k*). Furthermore*

$$c_{k,t} = \sum_{s-p=t} c_{k-1,p}\, a_{4p,s}\, 7^{2p-r} \tag{3.24}$$

if k *is even and the summation is over all integers* $p \geqslant 1$ *and* $s \geqslant [(8p+6)/7] = r$; *and*

$$c_{k,t} = \sum_{s-p=t} c_{k-1,t}\, a_{4p+1,s}\, 7^{2p-r}, \tag{3.25}$$

if k *is odd; the summation is over all integers* $p \geqslant 1$ *and* $s \geqslant [(8p+8)/7] = r$.

Lemma 6. *If for an integer* $k \geqslant 1$, $24\mu_k \equiv 1 \pmod{7^k}$ *and* $0 < \mu_k < 7^k$, *then*

$$L_k(x) = x\prod(1-x^{\delta n})\sum_{n=0}^{\infty} p(7^k n + \mu_k)\, x^n$$

where $\delta = 1$ *if* k *is even and* $= 7$ *if* k *is odd.*

For any integer $a \neq 0$, let $v(a)$ denote the highest power of 7 that divides a. Let for any integer $k \geqslant 1$,

$$\lambda_k = \min_t v(c_{k,t}).$$

From (3.24) and (3.25) it follows that

$$\lambda_k \geqslant \lambda_{k-1}, \quad k \geqslant 2.$$

Now $2p - r \geqslant 1$ if $p > 1$ and by (3.22), $a_{4,2} \equiv 0 \pmod 7$. It follows from (3.24), therefore that

$$\lambda_{2k} \geqslant \lambda_{2k-1} + 1. \tag{3.26}$$

From (3.23), we get

$$a_{2m-1} \cdot 7^m \equiv a_{2m-2} \cdot 7^m \cdot 190 \pmod{7^{m+1}},$$

which gives

$$a_{2m-1} \equiv a_{2m-2} \pmod 7. \tag{3.34}$$

Combining (3.33) and (3.34) we get, since $a_1 = 1$,

$$a_k = 5^{[\frac{2}{k}]} \pmod 7. \tag{3.35}$$

This shows that the congruences in theorem 4 cannot be improved and that

$$\lambda_k = \tfrac{1}{2}[k+2]. \tag{3.36}$$

We have

Corollary 2. *If k is odd, then*

$$p(7^k n + \mu_k) \equiv 5^{[\frac{2}{k}]} \cdot 7^{\frac{1}{2}(k+1)}\, p_{17}(n) \pmod{7^{\frac{1}{2}(k+3)}},$$

and if k is even

$$p(7^k n + \mu_k) \equiv 5^{k/2}\, 7^{k/2+1}\, p_{23}(n) \pmod{7^{(k/2)+2}}.$$

4. Higher moduli

4.1. In [14] Ramanujan gave proofs of

$$p(11n + 6) \equiv 0 \pmod{11},$$

$$p(11^2 n + 116) \equiv 0 \pmod{11^2}.$$

by using the functions P, Q, R. Actually the first one was published by Hardy, posthumously in [12]. Rushforth [19] completed the proof of the second using the ideas and methods of Ramanujan in [14]. In none of his unpublished manuscripts is there any mention of a method of attack for the modulus 11^n, $n \geqslant 3$. Even in his letter to Hardy he stops short of 11^2. As is well-known the conjecture for 11^n, $n \geqslant 1$ was proved completely by A O L Atkin [3].

However Ramanujan considered in [14] congruences to higher prime moduli. We shall consider the case of the modulus 13 since it has some similarity with 5 and 7.

Let us put

$$\varphi_{13}(x) = \varphi(x) = x^7 \frac{f(x^{169})}{f(x)}, \quad g_{13}(x) = g(x) = x \frac{f^2(x^{13})}{f^2(x)}.$$

Then if the operator U is defined as

$$U A(x) = \frac{1}{13} \sum_{\rho} A(\rho x^{1/13}),$$

where ρ runs through all the 13th roots of unity, then

$$U\varphi(x) = x \prod (1 - x^{13n}) \sum_{m=0}^{\infty} p(13m + 6)\, x^m. \tag{4.11}$$

Zuckermann proved [23], that

$$x\Pi(1-x^{13m})\sum_{m=0}^{\infty} p(13m+6)x^m$$

$$= 11g(x) + 13.36g^2(x) + 13^2 . 38g^3(x) + 13^3 .20g^4(x)$$

$$+ 13^4 . 6g^5(x) + 13^5 g^6(x) + 13^5 g^7(x). \qquad (4.12)$$

It seems *certain* that Ramanujan had proved this identity. On top of one of the three pages in the 'Lost' Note book [16] Ramanujan writes down, without comment, the numbers

$$11,\ 13.36,\ 13^2.38,\ 13^3.20,\ 13^4.6,\ 13^5,\ 13^5.$$

One can obtain this identity directly by computation or from the modular equation of degree 13, since $\varphi(\rho x^{1/13})$ are the roots of the modular equation.

From (4.12) we get at once, Ramanujan's congruence

$$\sum_{m=0}^{\infty} p(13m+6)x^{m+1} \equiv 11 . x\frac{f(x^{13})}{f^2(x)} \equiv 11\, x\Pi(1-x^n)^{11} \pmod{13}. \qquad (4.13)$$

From (4.12) we have

$$x\Pi(1-x^{13m})\sum_{m=0}^{\infty} p(13m+6)x^m \equiv 11\, g(x) \pmod{13}. \qquad (4.14)$$

If we apply the operator U to both sides of (4.14) we get, for the left side

$$x\Pi(1-x^m)\sum_{m=0}^{\infty} p(13^2 m + 13.12 + 6)x^m.$$

For the right side we have

$$Ug(x) \equiv U(x\Pi(1-x^m)^{24}) \equiv U\left(\sum_1^{\infty} \tau(m)x^m\right) \pmod{13}.$$

or
$$U(g(x)) \equiv \sum_1^m \tau(13m)x^m \pmod{13},$$

where $\tau(m)$ is defined in (3.31'). Ramanujan's results [17] give

$$\tau(13m) \equiv \tau(13) . \tau(m) \pmod{13}.$$

Since $\tau(13) \equiv 8 \pmod{13}$, we get

$$U(g(x)) \equiv 8\sum_1^{\infty} \tau(m)x^m \equiv 8g(x) \pmod{13}.$$

Therefore we get

$$x\Pi(1-x^m)\sum_0^{\infty} p(13^2 m + 162)x^m \equiv 10\, g(x) \pmod{13}. \qquad (4.15)$$

From (4.14) and (4.15) we obtain the Ramanujan congruences

$$\left.\begin{aligned} \sum_0^\infty p(13m+6)\,x^m &\equiv 11\,\Pi(1-x^m)^{11} \\ \sum_0^\infty p(13^2m+162)\,x^m &\equiv 10\,\Pi(1-x^m)^{23} \end{aligned}\right\} \pmod{13}. \tag{4.16}$$

If we are to continue further with powers of 13 on the left of (4.16), it seems necessary, following the cases 5 and 7, to obtain

$$U_{13}\,g(x) \quad \text{and} \quad U_{13}[\varphi(x)\,g(x)],$$

in powers of $g(x)$ with integers as coefficients. $U(g(x))$ is already given by Ramanujan's method. In order to determine $U(\varphi(x)\,g(x))$ one needs to study

$$U[\varphi(x)^3] = \frac{1}{13}\sum_\rho [\varphi(\rho x^{1/13})]^3, \tag{4.17}$$

ρ running through all the 13th roots of unity. If one knew the modular equation of degree 13 one could write (4.17) as a polynomial in $g(x)$ and know its coefficients. The modular equation of degree 13 is given by O'Brien [4]. It is, however, possible to prove, by direct computation, that

$$U(\varphi(x)\,g(x)) \equiv 4g(x) \pmod{13}.$$

If we write 1/24 in the 13-adic form

$$\frac{1}{24} = 6 + 12.13 + 5.13^2 + 12.13^3 + 5.13^4 + \ldots$$

so that

$$\mu_k = \begin{cases} \dfrac{11.13^{2m+1}+1}{24}, & k = 2m+1 \\ \dfrac{23.13^{2m}+1}{24}, & k = 2m. \end{cases}$$

then by the methods used for the moduli 5^α and 7^β, we get

$$x\Pi(1-x^{13m})\sum_0^\infty p(13^k m+\mu_k)\,x^m \equiv a_k\,g(x) \pmod{13},$$

if k is odd and

$$x\Pi(1-x^m)\sum_0^\infty p(13^k+\mu_k)\,x^m \equiv a_k\,g(x) \pmod{13}, \tag{4.18}$$

if k is even, a_k being an integer depending on k.

Using the congruences for $U[g(x)]$ and $U[\varphi(x)\,g(x)]$ we obtain

$$a_k \equiv \begin{cases} -2^{\frac{1}{2}(5k-3)} \pmod{13}, & k \text{ odd}, \\ -2^{\frac{1}{2}(5k-2)} \pmod{13}, & k \text{ even}. \end{cases}$$

Inserting in (4.18) the value of μ_k, we obtain finally the Ramanujan congruences

$$p\left(\frac{13^k(12n-1)+1}{24}\right) \equiv \begin{cases} -2^{\frac{1}{2}(5k-3)}\, p_{11}\left(\frac{n-1}{2}\right), & k, n \text{ odd}, \\ -2^{\frac{1}{2}(5k-2)}\, p_{23}\left(\frac{n-2}{2}\right), & k, n \text{ even}, \end{cases} \tag{4.19}$$

the congruences being to the modulus 13.

4.2. Morris Newman [8] obtained (4.19) independently by using a result [7] which he obtained by function-theoretic methods. Newman proved, what also follows from the previous section,

$$\left.\begin{aligned} p_{11}(13n+12) &\equiv 8p_{23}(n) \\ p_{23}(13n+5) &\equiv 4p_{11}(n) \end{aligned}\right\} \pmod{13}. \tag{4.21}$$

If therefore we can find n such that $p_{11}(n)$ or $p_{23}(n)$ is divisible by 13, then from (4.19) we would obtain some new congruences for $p(n)$ modulo 13. Newman [9] has tabulated the values of $p_{11}(n)$ upto $n = 700$. From the table we see that for

$$t = 17, 27, 57, 69, 95 \ldots$$

$$p_{11}(t) \equiv 0 \pmod{13}. \tag{4.22}$$

From (4.19) therefore we have

$$p\left(\frac{13^k m+1}{24}\right) \equiv 0 \pmod{13}. \tag{4.23}$$

for $m = 419$, 659, 1379, 1667 and 2291 and all odd k.

One can obtain a similar result for k and m even. The two results in (4.19) are in a sense equivalent in view of (4.21). All these are subsumed in Newman's work [8].

5. Remarks

5.1. Although Ramanujan says, in his letter to Hardy, that he had considered "more or less exhaustively about the congruency of $p(n)$ and in general that of $p_r(n)$ where

$$\sum_0^\infty p_r(n)\, x^n = 1/\{(1-x)(1-x^2)(1-x^3)\ldots\}^r \ldots",$$

there is not much about $p_r(n)$ in his unpublished manuscripts which are now available. In a page in the "Lost" note book [16], Ramanujan proves that

$$p_4(5n-1) \equiv 0 \pmod{5}, \tag{5.11}$$

where we have used, as in (2.59)

$$\sum_0^\infty p_r(n)\, x^n = \Pi(1-x^n)^r,$$

$r \neq 0$ is an integer (In Ramanujan's notation it will be $p_{-r}(n)$).

Ramanujan proves (5.11) by observing that

$$x\Pi(1-x^n)^4 = x(1-3x+5x^3-7x^6+\ldots)(1-x-x^2+\ldots)$$

$$= \sum_{\mu=0}^{\infty} \sum_{\nu=0}^{\infty} (-1)^{\mu+\nu}(2\nu+1)\, x^{1+\frac{1}{2}\mu(\mu+1)+\frac{1}{2}\nu(3\nu\pm1)}.$$

Now $1 + \frac{1}{2}\mu(\mu+1) + \frac{1}{2}\nu(3\nu+1) \equiv 0 \pmod 5$,

gives $(2\mu+1)^2 + 2(\nu+1)^2 \equiv 0 \pmod 5$,

which leads to $2\mu + 1 \equiv \nu + 1 \equiv 0 \pmod 5$ and hence to (5.11).

Ramanujan goes on to say "precisely in the same way we can show that

$$p_{-4}\left(np - \frac{p+1}{6}\right) \equiv 0 \pmod p,$$

where p is a prime of the form $6\lambda - 1$ and hence that

$$p_{sp-4}\left(np - \frac{p+1}{6}\right) \equiv 0 \pmod p.$$

For example

$$p_6(5n-1) \equiv 0 \pmod 5,\ p_7(11n-2) \equiv 0 \pmod{11}".$$

Ramanujan's assertions follow in the same way as above on observing that for a prime $p \equiv -1 \pmod 6$, the congruence $a^2 + 3b^2 \equiv 0 \pmod p$ holds only if $a \equiv b \equiv 0 \pmod p$.

There is one more page in this manuscript [16], where he writes

"If $\Pi(1-x^n)^5 = \sum_0^{\infty} q(n)\, x^n$

then $q(0) + q(5)\,x + q(10)\,x^2 + \ldots = \dfrac{\Pi(1-x^n)^6}{\Pi(1-x^{5n})}$.

If $\Pi(1-x^n)^7 = \sum_0^{\infty} q(n)\, x^n$, then

$$q(0) + q(7)\,x + q(14)\,x^2 + \ldots$$

$$= \frac{\Pi(1-x^n)^8}{\Pi(1-x^{7n})} + 49\,x\Pi(1-x^n)^4\,\Pi(1-x^{7n})^3 \ldots".$$

These results are stated without proof.

These two identities are also to be found in the paper of Morris Newman ([7], p. 320) where he says that D H Lehmer already knew them. Newman deduces these, and other such identities, as consequences of a general result proved by function-theoretic methods. They can be obtained by elementary methods which Ramanujan, perhaps, used.

Let us put

$$S_{-k} = \frac{1}{5} \sum_{\rho} [\varphi(\rho x^{1/5})]^{-k}$$

where the sum runs through all the 5th roots of unity and $\varphi(x) = \varphi_5(x)$ defined in (2.14). One obtains easily from the modular equation (2.15) that

$$S_{-1} = S_{-2} = -1,\ S_{-3} = S_{-4} = -5$$

$$S_{-5} = \frac{1}{g(x)}.$$

Therefore

$$\frac{1}{5}\sum_{\rho}(\varphi(\rho x^{1/5}))^{-5} = \frac{q(0) + q(5)\,x + q(10)\,x^2 + \ldots}{x\prod(1-x^{5x})^5}$$

$$= \frac{1}{x}\left(\frac{\prod(1-x^n)}{\prod(1-x^{5n})}\right)^6.$$

This gives Ramanujan's result. In a similar way the second identity can be proved.

On the same page of this manuscript [16], Ramanujan writes "n is the least positive integer such that $24n - 1$ is divisible by a positive integer k. Then

$$p(n + \nu k) - p(n)\,q(\nu) \tag{5.12}$$

is divisible by k for all positive integral values of ν where

$$\prod_r (1 - x^r)^{\frac{24n-1}{k}} = \sum_0^\infty q(\lambda)\,x^{\lambda}.$$

E.g.

$$p(4),\ p(9),\ p(14),\ \ldots \equiv 0 \pmod 5$$

$$p(6), p(17),\ p(28),\ \ldots \equiv 0 \pmod{11}$$

$$p(5), p(12),\ p(19),\ \ldots \equiv 0 \pmod 7$$

$$p(24) + 1,\ p(47) + 1,\ p(70),\ p(93),$$

$$p(116) - 1,\ p(139),\ p(162) - 1,\ p(185),\ \ldots \equiv 0 \pmod{23}$$

$$\ldots".$$

It is unfortunate that Ramanujan's assertion is incorrect. For instance the last two lines above are false. From the table of partitions (Collected Papers, p. 308) we have, for example

$$\left.\begin{aligned} p(24) + 1 &= 1575 + 1 = 1576 \\ p(47) + 1 &= 124754 + 1 = 124755 \\ p(93) \quad &= 82010177 \end{aligned}\right\} \not\equiv 0 \pmod{23}$$

Ramanujan's assertion is true for $k = 5^a,\ 11^a,\ a \geqslant 1$ and $k = 13$. It is also true for $k = 7,\ 7^2,\ 7^3,\ 7^4$.

$r \neq 0$ is an integer (In Ramanujan's notation it will be $p_{-r}(n)$).

Ramanujan proves (5.11) by observing that

$$x\Pi(1-x^n)^4 = x(1-3x+5x^3-7x^6+\ldots)(1-x-x^2+\ldots)$$

$$= \sum_{\mu=0}^{\infty} \sum_{\nu=0}^{\infty} (-1)^{\mu+\nu}(2+1)\, x^{1+\frac{1}{2}\mu(\mu+1)+\frac{1}{2}\nu(3\nu\pm1)}.$$

Now $1 + \frac{1}{2}\mu(\mu+1) + \frac{1}{2}\nu(3\nu+1) \equiv 0 \pmod 5$,

gives $(2\mu+1)^2 + 2(\nu+1)^2 \equiv 0 \pmod 5$,

which leads to $2\mu + 1 \equiv \nu + 1 \equiv 0 \pmod 5$ and hence to (5.11).

Ramanujan goes on to say " precisely in the same way we can show that

$$p_{-4}\left(np - \frac{p+1}{6}\right) \equiv 0 \pmod p,$$

where p is a prime of the form $6\lambda - 1$ and hence that

$$p_{sp-4}\left(np - \frac{p+1}{6}\right) \equiv 0 \pmod p.$$

For example

$$p_6(5n-1) \equiv 0 \pmod 5,\ p_7(11n-2) \equiv 0 \pmod{11}\ ".$$

Ramanujan's assertions follow in the same way as above on observing that for a prime $p \equiv -1 \pmod 6$, the congruence $a^2 + 3b^2 \equiv 0 \pmod p$ holds only if $a \equiv b \equiv 0 \pmod p$.

There is one more page in this manuscript [16], where he writes

" If $\Pi(1-x^n)^5 = \sum_0^{\infty} q(n)\, x^n$

then $q(0) + q(5)\, x + q(10)\, x^2 + \ldots = \dfrac{\Pi(1-x^n)^6}{\Pi(1-x^{5n})}$.

If $\Pi(1-x^n)^7 = \sum_0^{\infty} q(n)\, x^n$, then

$$q(0) + q(7)\, x + q(14)\, x^2 + \ldots$$

$$= \frac{\Pi(1-x^n)^8}{\Pi(1-x^{7n})} + 49\, x\Pi(1-x^n)^4\, \Pi(1-x^{7n})^3 \ldots ".$$

These results are stated without proof.

These two identities are also to be found in the paper of Morris Newman ([7], p. 320) where he says that D H Lehmer already knew them. Newman deduces these, and other such identities, as consequences of a general result proved by function-theoretic methods. They can be obtained by elementary methods which Ramanujan, perhaps, used.

Let us put

$$S_{-k} = \frac{1}{5} \sum_{\rho} [\varphi(\rho x^{1/5})]^{-k}$$

where the sum runs through all the 5th roots of unity and $\varphi(x) = \varphi_5(x)$ defined in (2.14). One obtains easily from the modular equation (2.15) that

$$S_{-1} = S_{-2} = -1, \; S_{-3} = S_{-4} = -5$$

$$S_{-5} = \frac{1}{g(x)}.$$

Therefore

$$\frac{1}{5} \sum_{\rho} (\varphi(\rho x^{1/5}))^{-5} = \frac{q(0) + q(5)\,x + q(10)\,x^2 + \dots}{x\Pi(1 - x^{5n})^5}$$

$$= \frac{1}{x}\left(\frac{\Pi(1 - x^n)}{\Pi(1 - x^{5n})}\right)^6.$$

This gives Ramanujan's result. In a similar way the second identity can be proved.

On the same page of this manuscript [16], Ramanujan writes "n is the least positive integer such that $24n - 1$ is divisible by a positive integer k. Then

$$p(n + \nu k) - p(n)\,q(\nu) \tag{5.12}$$

is divisible by k for all positive integral values of ν where

$$\prod_r (1 - x^r)^{\frac{24n-1}{k}} = \sum_0^\infty q(\lambda)\,x^\lambda.$$

E.g.

$$p(4),\ p(9),\ p(14), \dots \equiv 0 \pmod 5$$

$$p(6), p(17),\ p(28), \dots \equiv 0 \pmod{11}$$

$$p(5), p(12),\ p(19), \dots \equiv 0 \pmod 7$$

$$p(24) + 1,\ p(47) + 1,\ p(70),\ p(93),$$

$$p(116) - 1,\ p(139),\ p(162) - 1,\ p(185), \dots \equiv 0 \pmod{23}$$

$$\dots".$$

It is unfortunate that Ramanujan's assertion is incorrect. For instance the last two lines above are false. From the table of partitions (Collected Papers, p. 308) we have, for example

$$\left.\begin{aligned} p(24) + 1 &= 1575 + 1 = 1576 \\ p(47) + 1 &= 124754 + 1 = 124755 \\ p(93) \quad &= 82010177 \end{aligned}\right\} \not\equiv 0 \pmod{23}$$

Ramanujan's assertion is true for $k = 5^a, 11^a$, $a \geqslant 1$ and $k = 13$. It is also true for $k = 7, 7^2, 7^3, 7^4$.

Appendix

Firtzroy House
16, Fritzroy Square
Monday

Dear Mr. Hardy,

Please tell Mr. Littlewood and Major Mac Mahon that I thanked them very much. Had it not been for your pains and their encouragement, I would be neither the Fellow of the one nor that of the other ...

I have considered more or less exhaustively about the congruency of $p(n)$ and in general that of $p_r(n)$ where

$$\Sigma\, p_r(n)\, x^n = 1/\{(1-x)(1-x^2)(1-x^3)\ldots\}^r$$

by four different methods. Each method has its own advantages. In the case of $p(n)$ the results are roughly these :

I method. Very special and very simple. Gives only $p(5n+4) \equiv 0 \pmod 5$

$p(7n+5) \equiv 0 \pmod 7$. This you are publishing now.

II ϑ-function method. This is general theoretically but becomes practically unworkable due to tediousness of calculation with the exception of few cases. I have simplified the proof very much and made it quite elementary as the first method. From this the divisibility by 5, 5^2, 5^3 and 7, 7^2, 7^3 is established quite easily. I have not considered the case of 11 due to tediousness. This gives many more results like (17) and (18) in the proof sheet.

III Dirichlet's series method. This is very general and practically workable. The conjectured products of D-series corresponding to

$\{(1-x)(1-x^2)(1-x^3)\ldots\}^n$ where $n = 2, 4, 6, 8, 12, 24$ were proved by Mr. Mordell. I have now conjectured analogous results for some more values of n, viz., 10, 14, 16, 18, 20, 22, 28, 30, 32, 36, 48. All these may be proved by Mr. Mordell. If we assume these results this method gives information about the divisibility of $p(n)$ by

5, 5^2, 7, 7^2, 11, 13, 17, 19, 23, 29, 31 and 37.

IV g_2, g_3 method. This is very general. This gives all the previous results, except the case of 5^3, 7^3 and the divisibility by 11^2 as well.

Thus the divisibility by $5^a\, 7^b\, 11^c$ when $a = 0, 1, 2, 3$, $b = 0, 1, 2, 3$, $c = 0, 1, 2$ amounting to $4 \times 4 \times 3 - 1$ or 47 cases of the conjectured theorem are proved.

Ever yours
S. Ramanujan.

References

[1] Andrews G E 1979 An introduction to Ramanujan's "Lost" Note book *Am. Math. Monthly* 1979 **86** 89–108.

[2] Atkin A O L 1968 Ramanujan congruences for $p_{-k}(n)$; *Can J. Math.* **10** 67–78

[3] Atkin A O L 1967 Proof of a conjecture of Ramanujan; *Glasgow. Math. J.* **8** 14–32

[4] Atkin A O L and O'Brien J N 1967 Some properties of $p(n)$ and $c(n)$ modulo powers of 13 ; *Trans. Am. Math. Soc.* **126** 442–459

[5] Birch B J 1975 A lookback at Ramanujan's notebooks ; *Proc. Camb. Philos. Soc.* **78** 73–79

[6] Hardy G H and Wright E M 1976 *An introduction to the theory of numbers* (Oxford :

[7] Newman M 1952 Remarks on some modular identities ; *Trans. Am. Math. Soc.* **73** 313–320

[8] Newman M 1957 Congruences for the coefficients of modular forms and some new congruences for the partition function ; *Can. J. Math.* **9** 549–552

[9] Newman M 1956 A table of the coefficients of the powers of $\eta(\tau)$; *Indagationes. Math.* **18** 204–216

[10] Ramanujan S 1919 Some properties of $p(n)$, the number of partitions of n ; *Proc. Camb. Phil. Soc.* **19** 207–210 (Collected Papers 210–213)

[11] Ramanujan S 1920 Congruence properties of partitions ; *Proc London. Math. Soc.* **2** (18) (Collected Papers, 230)

[12] Ramanujan S 1921 Congruence properties of partitions ; *Math. Zeit.* **9** 147–153 (Collected Papers 232–238)

[13] Ramanujan S 1959 *Note books of Srinivasa Ramanujan,* Tata Institute of Fundamental Research

[14] Ramanujan S 1920 Properties of $p(n)$ and $\tau(n)$ defined by the equations. Unpublished manuscript, Trinity College Library

[15] Ramanujan S 1980 *Notes copied by G. N. Watson* (Oxford : Mathematical Library)

[16] Ramanujan S 1978 *A "Lost" Notebook* (Cambridge : Trinity College)

[17] Ramanujan S 1916 On certain arithmetical functions. *Trans. Camb. Phil. Soc.* **22** 159–184 (Collected Papers 136–162)

[18] Rankin R A 1976 *Ramanujan's unpublished work on congruences. Modular functions of one variable. VI* (Bonn : Springer-Verlag Lecture Notes) **601** 3–14

[19] Rushforth J M 1952 Congruence properties of the partition function and associated functions ; *Proc. Camb Philos. Soc.* **48** 402–413

[20] Stanley G K 1928 Two assertions made by Ramanujan ; *J. London Math. Soc.* **3** 232–237

[21] Watson G N 1935 Über Ramanujansche Kongruenzeigenschaften der Zerfällungsanzahlen ; *Math. Zeit.* **39** 712–731

[22] Watson G N 1938 Ramanujans vermutung uber Zerfällungsanzahlen ; *Jour. fur. Math.* **179** 97–128

[23] Zuckerman H S 1939 Identities analogous to Ramanujan's identities involving the partition function ; *Duke. Math. Jour.* **5** 88–110

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 3, September 1980, pp. 159–170.

Diffraction of impulsive elastic waves by a fluid cylinder

B K RAJHANS and S K MISHRA*
Department of Physics and Mathematics, Indian School of Mines, Dhanbad 826 004, India
* Department of Mathematics, Patna University, Patna 800 005, India

MS received 11 July 1979

Abstract. We consider the diffraction of impulsive SV waves by a fluid circular cylinder. The cylinder is embedded in an unbounded isotropic homogeneous elastic medium and it is filled with some acoustic fluid. The line source, generating the incident pulse, is situated outside the cylinder parallel to its axis. We investigate the problem by the method of dual integral transformation as developed by Friedlander. The resulting integrals are evaluated approximately to obtain the short-time estimate of the motion near the wave front in the shadow zone of the elastic medium. We also interpret the approximate solution in terms of Keller's geometrical theory of diffraction.

Keywords. Elastic waves; diffraction; dual integral transformation; Keller's geometric theory; pulse propagation mode.

1. Introduction

The diffraction of two-dimensional elastic waves with a cylindrical obstacle in an unbounded medium has been considered in recent years. Gilbert and Knopoff [6] discussed the scattering of impulsive elastic waves by a rigid circular cylinder situated in a homogeneous isotropic elastic medium. Gilbert [5] considered the scattering of impulsive elastic waves by a smooth convex cylinder and obtained the formal solution of the problem using the technique of dual integral transformation developed by Friedlander [2]. An approximate evaluation of the solution was then obtained corresponding to the short-time behaviour of the scattered field near the wave-front both in the shadow as well as in the illuminated zone. Jha [7] also used the same technique to investigate the problem of diffraction of compressional waves by a fluid cylinder in a homogeneous medium.

In this paper, we investigate the diffraction of the impulsive SV pulses by a circular cylindrical obstacle filled with inviscid fluid material. The obstacle is supposed to be situated in an unbounded homogeneous isotropic elastic medium and the incident pulse is generated by a line source situated in the surrounding elastic medium at a finite distance parallel to the axis of the cylinder. We discuss the scattering of SV waves. It is well-known that lime stone formations usually exhibit a shear-wave arrival stronger than the compressional waves. Therefore

in such cases shear-waves are more prominent than the compressional waves [13]. We suppose that the velocities of P and SV waves outside the cylinder are α and β respectively and that of P waves inside the cylinder is α_0. To be specific, we assume $\alpha > \alpha_0 > \beta$. This assumption of the velocity distribution corresponds to the actual velocity distribution of elastic waves inside the earth and to the location of the source in the mantle and the outer core as the obstacle [1]. The present investigation therefore throws some light on the scattering of SV waves by the core of the earth. We also suppose the density of the medium outside the cylinder is ρ and that inside the cylinder is ρ' where $\rho > \rho'$.

2. Formulation of the problem

Let the axis of the cylinder be taken as the z-axis and let an (r, θ) co-ordinate system be located in the (x, y) plane with $\theta = 0$, $r = r_0 (> a)$ corresponding to the location of the line source which is parallel to the axis of the cylinder. The equation of the cylinder is $r = a$.

We define the elastic velocity potentials ϕ_0, ϕ and ψ corresponding to the wave equations inside and outside the cylinder respectively. Since z-axis is taken along the axis of the cylinder, the state of the media is fully determined if the velocity potentials ϕ_0, ϕ and ψ are obtained as a function of r, θ and t. It is thus required to find out the velocity potentials ϕ_0, ϕ and ψ as a function of r, θ and t which satisfy the wave equations

$$\frac{1}{\beta^2}\frac{\partial^2\psi}{\partial t^2} - \nabla^2\psi = \frac{2\pi}{r}\delta(r - r_0)\,\delta(t)\,\delta(\theta), \quad (r \geqslant a), \tag{1}$$

$$\frac{1}{\alpha^2}\frac{\partial^2\phi}{\partial t^2} - \nabla^2\phi = 0, \quad (r \geqslant a), \tag{2}$$

$$\frac{1}{\alpha_0^2}\frac{\partial^2\phi_0}{\partial t^2} - \nabla^2\phi_0 = 0, \quad (r \leqslant a), \tag{3}$$

where $\nabla^2 = \dfrac{\partial^2}{\partial r^2} + \dfrac{1}{r}\dfrac{\partial}{\partial r} + \dfrac{1}{r^2}\dfrac{\partial^2}{\partial\theta^2}$;

the initial conditions

$$\begin{aligned} \psi &= \partial\psi/\partial t = 0, \text{ when } t = 0 \text{ except at } r = r_0,\ \theta = 0, \\ \phi &= \partial\phi/\partial t = 0, \text{ when } t = 0, \\ \phi_0 &= \partial\phi_0/\partial t = 0, \text{ when } t = 0 \end{aligned} \tag{4}$$

and the boundary conditions

$$\begin{aligned} &[T_{r\theta}]_{r=a+0} = 0, \\ &[T_{rr}]_{r=a+0} = [T_{rr}]_{r=a-0}, \\ &[u_r]_{r=a+0} = [u_r]_{r=a-0}, \end{aligned} \tag{5}$$

where $T_{r\theta} = \nu\left\{\dfrac{1}{r}\dfrac{\partial u_r}{\partial\theta} + \dfrac{\partial u_\theta}{\partial r} - \dfrac{u_\theta}{r}\right\}$,

$$T_{rr} = \lambda\frac{1}{\alpha^2}\frac{\partial^2\phi}{\partial t^2} + 2\nu\frac{\partial u_r}{\partial r},$$

$$u_r = \frac{\partial \phi}{\partial r} + \frac{1}{r}\frac{\partial \psi}{\partial \theta},$$

$$u_\theta = \frac{1}{r}\frac{\partial \phi}{\partial \theta} - \frac{\partial \psi}{\partial r},$$

Here λ and ν are Lame's parameters and the symbol δ stands for Dirac delta function.

3. The formal solution

Let us define the Laplace transform $\bar{\psi}(r, \theta, s)$ of $\psi(r, \theta, t)$ by

$$\bar{\psi}(r, \theta, s) = \int_0^\infty \psi(r, \theta, t) \exp(-st)\, dt, \tag{6}$$

where s is the transform variable. Again we denote the Fourier transform ψ^* of $\bar{\psi}$ by

$$\psi^*(r, \mu, s) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \bar{\psi}(r, \theta, s) \exp(-i\mu\theta)\, d\theta. \tag{7}$$

Applying these transformations to (1), (2) and (3) we get

$$\frac{d^2\psi^*}{dr^2} + \frac{1}{r}\frac{d\psi^*}{dr} - \left(\frac{s^2}{\beta^2} + \frac{\mu^2}{r^2}\right)\psi^* = -\frac{(2\pi)^{1/2}}{r}\delta(r - r_0), \quad (r \geqslant a) \tag{8}$$

$$\frac{d^2\phi^*}{dr^2} + \frac{1}{r}\frac{d\phi^*}{dr} - \left(\frac{s^2}{\alpha^2} + \frac{\mu^2}{r^2}\right)\phi^* = 0, \quad (r \geqslant a), \tag{9}$$

and

$$\frac{d^2\phi_0^*}{dr^2} + \frac{1}{r}\frac{d\phi_0^*}{dr} - \left(\frac{s^2}{\alpha_0^2} + \frac{\mu^2}{r^2}\right)\phi_0^* = 0, \quad (r \leqslant a). \tag{10}$$

Also the Laplace–Fourier transformed boundary conditions are given by

$$[T^*_{r\theta}]_{r=a+0} = 0,$$

$$[T^*_{rr}]_{r=a+0} = [T^*_{rr}]_{r=a-0}, \tag{11}$$

and $$[U^*_r]_{r=a+0} = [U^*_r]_{r=a-0}.$$

It may be assumed that ψ^* is continuous at $r = r_0$.
Then (8) is equivalent to

$$\frac{d^2\psi^*}{dr^2} + \frac{1}{r}\frac{d\psi^*}{dr} - \left(\frac{s^2}{\beta^2} + \frac{\mu^2}{r^2}\right)\psi^* = 0, \quad (r \geqslant a), \tag{12}$$

and

$$[\psi^*]_{r_0-0}^{r_0+0} = 0, \quad [d\psi^*/dr]_{r_0-0}^{r_0+0} = -(2\pi)^{1/2}/r_0. \tag{13}$$

After a little mathematical calculation and Fourier inversion one finds that the Laplace transforms of the solution are given by

$$\bar{\psi}(r,\theta,s) = \int_{-\infty}^{\infty} I_{|\mu|}(sr/\beta)\, K_\mu(sr_0/\beta) \exp(i\mu\theta)\, d\mu$$

$$+ \int_0^{\infty} K_\mu(sr/\beta)\, K_\mu(sr_0/\beta)\, \frac{L(\mu)}{M(\mu)} \exp(i\mu\theta)\, d\mu$$

$$+ \int_{-\infty}^{0} K_\mu(sr/\beta)\, K_\mu(sr_0/\beta_0)\, \frac{L(-\mu)}{M(-\mu)} \exp(i\mu\theta)\, d\mu.$$

$$(r_0 \geqslant r \geqslant a) \qquad (14)$$

$$\bar{\phi}(r,\theta,s) = \int_0^{\infty} K_\mu(sr/\alpha)\, K_\mu(sr_0/\beta)\, \frac{N(\mu)}{M(\mu)} \exp(i\mu\theta)\, d\mu$$

$$+ \int_{-\infty}^{0} K_\mu(sr/\alpha)\, K_\mu(sr_0/\beta)\, \frac{N(-\mu)}{M(-\mu)} \exp(i\mu\theta)\, d\mu$$

$$(r \geqslant a) \qquad (15)$$

where $L(\mu) = \frac{2\rho s^4}{a\beta\alpha_0} I'_\mu(sa/\alpha_0)\, I'_\mu(sa/\beta)\, K_\mu(sa/\alpha) - \frac{4\rho s^3 \beta}{a^2 \alpha\alpha_0} I'_\mu(sa/\alpha_0)$

$$\times I'_\mu(sa/\beta)\, K'_\mu(sa/\alpha) - \frac{2\rho' s^4}{a\alpha\beta} I_\mu(sa/\alpha_0)\, I'_\mu(sa/\beta)\, K'_\mu(sa/\alpha)$$

$$- \frac{\rho s\beta^2}{\alpha_0}\left(\frac{s^2}{\beta^2} + \frac{2\mu^2}{a^2}\right) I'_\mu(sa/\alpha_0)\, I_\mu(sa/\beta)\, K_\mu(sa/\alpha)$$

$$+ \frac{2\rho s^4}{a\alpha\alpha_0} I'_\mu(sa/\alpha_0)\, I_\mu(sa/\beta)\, K'_\mu(sa/\alpha) + \frac{\rho' s^5}{a\beta^2} I_\mu(sa/\alpha_0)$$

$$\times I_\mu(sa/\beta)\, K'_\mu(sa/\alpha) + \frac{4\rho\mu^2 s^3 \beta}{a^2 \alpha\alpha_0} I'_\mu(sa/\alpha_0)\, I'_\mu(sa/\beta)\, K'_\mu(sa/\alpha)$$

$$+ \frac{4\rho s\mu^2 \beta^2}{a^4 \alpha_0} I'_\mu(sa/\alpha_0)\, I_\mu(sa/\beta)\, K_\mu(sa/\alpha)$$

$$+ \frac{2\rho' \mu^2 s^2}{a^3} I_\mu(sa/\alpha_0)\, I_\mu(sa/\beta)\, K_\mu(sa/\alpha),$$

$$M(\mu) = \frac{4\rho s^3 \beta}{a^2 \alpha\alpha_0} I'_\mu(sa/\alpha_0)\, K'_\mu(sa/\alpha)\, K'_\mu(sa/\beta) - \frac{4\rho\beta\mu^2 s^3}{a^2 \alpha\alpha_0}$$

$$\times I'_\mu(sa/\alpha_0)\, K'_\mu(sa/\alpha)\, K'_\mu(sa/\beta) - \frac{4\rho s\mu^2 \beta^2}{a^4 \alpha_0} I'_\mu(sa/\alpha_0)$$

$$\times K_\mu(sa/\alpha)\, K_\mu(sa/\beta) - \frac{2\rho' \mu^2 s^2}{a^3} I_\mu(sa/\alpha_0)\, K_\mu(sa/\alpha)\, K_\mu(sa/\beta)$$

$$- \frac{2\rho s^4}{a\beta\alpha_0} I'_\mu(sa/\alpha_0)\, K_\mu(sa/\alpha)\, K'_\mu(sa/\beta) + \frac{2\rho' s^4}{a\alpha\beta} I_\mu(sa/\alpha_0)$$

$$\times K'_\mu(sa/\alpha)\,K'_\mu(sa/\beta)+\frac{\rho s\beta^2}{\alpha_0}\left(\frac{s^2}{\beta^2}+\frac{2\mu^2}{a^2}\right)^2 I'_\mu(sa/\alpha_0)$$

$$\times K_\mu(sa/\alpha)\,K_\mu(sa/\beta)-\frac{2\rho s^4}{a\alpha\alpha_0}\,I'_\mu(sa/\alpha_0)\,K'_\mu(sa/\alpha)\,K_\mu(sa/\beta)$$

$$-\frac{\rho' s^5}{a\beta^2}\,I_\mu(sa/\alpha_0)\,K'_\mu(sa/\alpha)\,K_\mu(sa/\beta),$$

$$N(\mu)=\frac{4i\mu\rho s\beta^2}{a^4\alpha_0}\,I'_\mu(sa/\alpha_0)+\frac{2i\mu\rho' s^2}{a^3}\,I_\mu(sa/\alpha_0)$$

$$-\frac{2i\mu\rho s\beta^2}{a^2\alpha_0}\left(\frac{s^2}{\beta^2}+\frac{2\mu^2}{a^2}\right)I'_\mu(sa/\alpha_0), \tag{16}$$

The expression for $\bar{\psi}(r,\theta,s)$ in $r\geqslant r_0$ is obtained from (14) by interchanging r and r_0. We are only interested in the solution outside the obstacle.

It is clear that (14) and (15) give the integral representation of Laplace transform of the formal solution. The time solution can be obtained on performing Laplace inversion. But it is difficult to evaluate the integrals (14) and (15) as they stand. However, if one is interested in short-time behaviour of pulses, these integrals can be evaluated approximately for large, positive and real s. For this purpose we use the method of residues. The solution is then, represented by a series of residues evaluated at the poles of the integrand which are given by the zeros of $M(\mu)$ and $M(-\mu)$ respectively.

4. Zeros of M(μ) and M(− μ)

Now we proceed to investigate the zeros of $M(\mu)$ in the complex μ-plane. The investigation of the zeros of $M(-\mu)$ is similar. If $s\gg|\mu|$, we note that $M(\mu)$ has no zeros. This can be easily verified by using standard approximation for modified Bessel functions of large arguments [12]. When $|\mu|$ and s are both large but μ is not pure imaginary, we arrive at the same conclusion if we use the corresponding asymptotic approximations [11]. Thus we are left with the case when $|\mu|$ and s are both large and μ is pure imaginary. The appropriate approximations to be used in this case are given by Mishra [9]. Now we can investigate the zeros of $M(\mu)$ when μ is pure imaginary and both $|\mu|$ and s are large. Let $\mu=iv$ where v is real and positive. The following situations arise

(a) $v<\frac{sa}{\alpha}<\frac{sa}{\alpha_0}<\frac{sa}{\beta}$, (b) $v\sim\frac{sa}{\alpha}<\frac{sa}{\alpha_0}<\frac{sa}{\beta}$,

(c) $\frac{sa}{\alpha}<\frac{sa}{\alpha_0}<\frac{sa}{\beta}\sim v$, (d) $\frac{sa}{\alpha}<\frac{sa}{\alpha_0}<\frac{sa}{\beta}<v$,

(e) $\frac{sa}{\alpha}<v<\frac{sa}{\alpha_0}<\frac{sa}{\beta}$, (f) $\frac{sa}{\alpha}<\frac{sa}{\alpha_0}<v<\frac{sa}{\beta}$, and

(g) $\frac{sa}{\alpha}<\frac{sa}{\alpha_0}\sim v<\frac{sa}{\beta}$.

If the appropriate approximations as given by Mishra [9] are used, we find that in cases (a), (d), (e), (f) and (g) there are no zeros of $M(\mu)$. We have not given

here the detailed expressions as they are quite unwieldy. In cases (b) and (c), $M(\mu)$ possesses an infinite set of zeros. These are the zeros of $f(x)$ where $f(x)$ is defined by

$$f(x) = 3(2)^{-1/6}\,\mathrm{Ai}(-2^{1/3}x)$$

$$F(x) = -(3)^{1/2}(2)^{-1/6}\,\mathrm{Bi}(-2^{1/3}x),$$

$$v = s + xs^{1/3} + o(s^{-1/3}). \tag{17}$$

It is well-known that the zeros of Airy functions are all real and negative. Let the zeros of Ai $(-2^{1/3}x)$ be x_j, $j = 1, 2, 3\cdots$ so that x_j is real and positive. Then by the third equation of (17), the zeros of $M(\mu)$ may be written as

$$v_j = \frac{sa}{a} + x_j(sa/a)^{1/3} + o(s^{-1/3}). \tag{18}$$

By similar arguments the zeros of $M(\mu)$ in case (c) may be written as

$$v_j = \frac{sa}{\beta} + x_j(sa/\beta)^{1/3} + o(s^{-1/3}). \tag{19}$$

We recall that we have put $\mu = iv$, v being real and positive. Hence the zeros of $M(\mu)$ can be written as

$$\mu = iv_j, \tag{20}$$

where v_j is given by (18) and (19). Following the arguments of Mishra [9], it can be shown that $M(\mu)$ and $M(-\mu)$ have the same set of zeros.

5. Pulse propagation modes

Now the integrals appearing in (14) and (15) can be evaluated. First we consider (14). The first integral in (14) has no contribution as its integrand is regular in the whole μ-plane. In order to evaluate the integral from 0 to ∞, we close the contour in the first quadrant of the complex μ-plane and introduce indentations round the poles given by (18), (19) and (20) on the imaginary axis (figure 1). Similarly, the contour is closed in the second quadrant to evaluate the integral from $-\infty$ to 0. It can be shown that the contribution to each integral from the quadrant part of the contour tends to zero as its radius tends to infinity. The same procedure is applied while evaluating (15). Thus we find that $\bar{\psi}$ and $\bar{\phi}$ are formally given by

$$\bar{\psi}(r, \theta, s) = \sum_{j=1}^{\infty} \bar{\psi}_j + B_1, \tag{21}$$

where

$$\bar{\psi}_j = -\frac{\pi L(iv_j)}{\left[\frac{\partial}{\partial v} M(iv)\right]_{v=v_j}} K_{iv_j}(sr/\beta)\, K_{iv_j}(sr_0/\beta) \exp\{-v_j \mid \theta \mid\}$$

$$-\frac{\pi L(-iv_j)}{\left[\frac{\partial}{\partial v} M(-iv)\right]_{v=v_j}} K_{iv_j}(sr/\beta)\, K_{iv_j}(sr_0/\beta) \exp\{-v_j \mid \theta \mid\},$$

$$(r_0 \geqslant r \geqslant a)$$

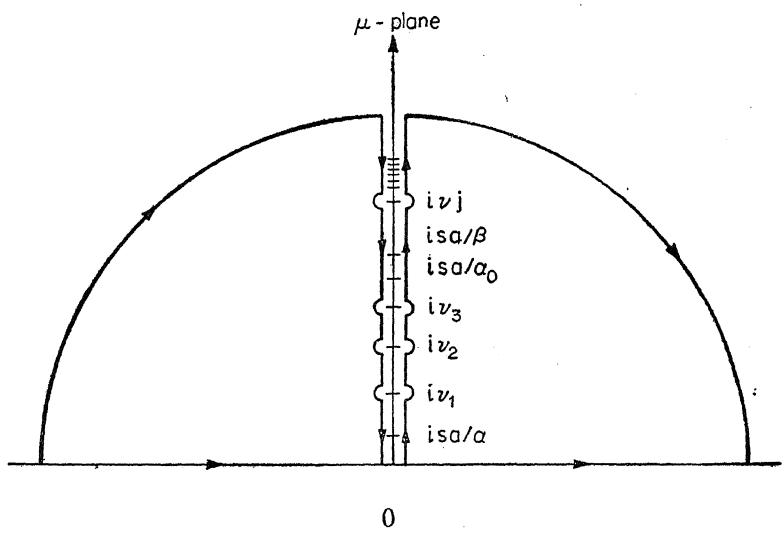

Figure 1. Zeros of the function $M(\mu)$.

and $$\bar{\phi}(r, \theta, s) = \sum_{j=1}^{\infty} \bar{\phi}_j + B_2 \tag{22}$$

where $$\bar{\phi}_j = -\frac{\pi N(iv_j)}{\left[\frac{\partial}{\partial \nu} M(i\nu)\right]_{\nu=\nu j}} K_{i\nu j}(sr/\alpha)\, K_{i\nu j}(sr_0/\beta) \exp\{-\nu_j \mid \theta \mid\}$$

$$-\frac{\pi N(-iv_j)}{\left[\frac{\partial}{\partial \nu} M(-i\nu)\right]_{\nu=\nu_j}} K_{i\nu j}(sr/\alpha)\, K_{i\nu j}(sr_0/\beta) \exp\{-\nu_j \mid \theta \mid\},$$

$$(r \geqslant a)$$

where B_1 and B_2 stand for the line integrals along the imaginary axis. If we follow the arguments of Mishra [9, 10], we can say that the contribution due to these line integrals may be ignored. Each $\bar{\psi}_j$ and $\bar{\phi}_j$ is the Laplace transform of a pulse which we will call a 'pulse propagation mode' following Friedlander [4]. It is not possible to evaluate the inverse Laplace transform of (21) and (22) as they stand. However, we are interested in the behaviour of the pulses in the wake of the onset. We know that this is correlated with the behaviour of its Laplace transform for large s. It has been assumed that s is real and positive. Therefore $\bar{\psi}(r, \theta, s)$ and $\bar{\phi}(r, \theta, s)$ will be evaluated for large, real and positive s. To this end, we will use the various approximations for modified Bessel functions of imaginary order which have been given by Mishra [9]

First, the approximations for $\bar{\psi}_j$ will be found. From (19) we have

$$\bar{\nu}_j \sim sa/\beta + x_j (sa/\beta)^{1/3}, \tag{23}$$

where x_j is the jth zero of $f(x)$. We can write

$$f(x_j) = 0 \quad (j = 1, 2, 3 \cdots), \tag{24}$$

where by (17), we have

$$f(x_j) = (3/2^{1/6})\,\mathrm{Ai}\,(-2^{1/3}x_j). \tag{25}$$

It is clear from (23) that the operator

$$[\partial/\partial\nu]_{\nu=\nu_j}$$

can be replaced by

$$(sa/\beta)^{-1/3}\ [\partial/\partial x]_{s=s_j}$$

From (17), we see that

$$f''(x) = (3/2^{1/6})\,2^{2/3}\,\mathrm{Ai}''\,(-2^{1/3}x). \tag{26}$$

But we know that the Airy functions are solutions of the differential equation

$$d^2w/dz^2 = zw. \tag{27}$$

Therefore we have $f''(x) = -2xf(x)$. From (24) and (28), we get (28)

$$f''(x_j) = 0. \tag{29}$$

It may be recalled that $\alpha > \alpha_0 > \beta$. Therefore one can put

$$\begin{aligned} \alpha &= n\beta\,(n > 1), \\ \alpha_0 &= n_1\beta\,(n_1 > 1), \\ \alpha &= n_2\alpha_0\,(n_2 > 1). \end{aligned} \tag{30}$$

Now if we use the appropriate approximations and make use of (23), we find that

$$\begin{aligned} M(i\nu) \sim -\frac{2^{1/2}\,\pi s^{1/3}}{3\alpha^{4/3}\,\beta^{5/3}}\Bigg[\frac{i\rho\alpha^{1/2}}{\alpha_0^{1/2}}\frac{(n_1^2-1)^{1/4}}{(n^2-1)^{1/4}}\sin\Bigg[\frac{\pi}{4}+\frac{sa}{\beta}\Bigg\{\cosh^{-1}n \\ -\left(1-\frac{1}{n^2}\right)^{1/2}\Bigg\}+x\,(sa/\beta)^{1/3}\cosh^{-1}n\Bigg]-\frac{\rho'\,\alpha_0^{\frac{1}{2}}}{\alpha^{1/2}}\frac{(n^2-1)^{1/4}}{(n_1^2-1)^{1/4}}, \\ \cos\Bigg[\frac{\pi}{4}+\frac{sa}{\beta}\Bigg\{\cosh^{-1}n-\left(1-\frac{1}{n^2}\right)^{1/2}\Bigg\}+x\left(\frac{sa}{\beta}\right)^{1/3}\cosh^{-1}n\Bigg] \\ f(x)\exp\Bigg[-\frac{sa}{\beta}\Bigg\{\frac{\pi}{2}+i\cosh^{-1}n_1-i\left(1-\frac{1}{n_1^2}\right)^{1/2}\Bigg\} \\ -x\,(sa/\beta)^{1/3}\Bigg\{\frac{\pi}{2}+i\cosh^{-1}n_1\Bigg\}-\frac{i\pi}{4}\Bigg]. \end{aligned} \tag{31}$$

We also have the following approximations for $\bar{\psi}_j$ when

$$\nu_j \sim (sa/\alpha) + x_j\,(sa/\alpha)^{1/3}.$$

In this case the operator $[\partial/\partial\nu]_{\nu=\nu_j}$ can be replaced by

$$(sa/\alpha)^{-1/3}\ [\partial/\partial x]_{s=s_j},$$

$$\left[\frac{\partial}{\partial v} M(iv)\right]_{v=\nu_j} \sim \frac{\pi\rho s^{10/3}\beta^2}{3\sqrt{2}\, a^{5/3}\alpha^{11/3}} \frac{(n^2-2)^2 (n_2^2-1)^{1/4}}{(n^2-1)^{1/4}} f'(x_j) \cdot$$

$$\exp\left[\frac{sa}{\alpha}\left\{(n_2^2-1)^{1/2} - (n^2-1)^{1/2} + \cos^{-1}\frac{1}{n} - \cos^{-1}\frac{1}{n_2} - \frac{\pi}{2}\right\}\right.$$

$$\left. + x_j (sa/\alpha)^{1/3}\left\{\cos^{-1}\frac{1}{n} - \cos^{-1}\frac{1}{n_2} - \frac{\pi}{2}\right\}\right]. \tag{32}$$

The approximations for $\bar{\phi}_j$ are given as follows; when

$$\nu_j \sim \frac{sa}{\alpha} + x_j (sa/\alpha)^{1/3}$$

$$\left[\frac{\partial}{\partial \nu} M(i\nu)\right]_{\nu=\nu_j} \sim \frac{\pi\rho s^{10/3}\beta^2}{3\sqrt{2}\, a^{5/3}\alpha^{11/3}} \frac{(n^2-2)^2 (n_2^2-1)^{1/4}}{(n^2-1)^{1/4}} f'(x_j) \cdot$$

$$\exp\left[\frac{sa}{\alpha}\left\{(n_2^2-1)^{1/2} - (n^2-1)^{1/2} + \cos^{-1}\frac{1}{n} - \cos^{-1}\frac{1}{n_2} - \frac{\pi}{2}\right\}\right.$$

$$\left. + x_j (sa/\alpha)^{1/3}\left\{\cos^{-1}\frac{1}{n} - \cos^{-1}\frac{1}{n_2} - \frac{\pi}{2}\right\}\right]. \tag{33}$$

We note that these approximations remain unchanged if ν is replaced by $-\nu$ or i is replaced by $-i$. Therefore, if we use the approximations for modified Bessel functions given by Mishra [9], we obtain from (21) and (22)

$$\psi_j^{(1)}(r,\theta,s) \sim -\sum_{j=1}^{\infty} \frac{\pi a^{4/3}}{s}\left[\frac{4\rho\beta^4 (n^2-1)^{1/2}(n_2^2-1)^{1/2} - \rho' a^4}{\rho\beta^4 (n_2^2-1)^{1/4}}\right] \cdot$$

$$\frac{1}{(n^2-2)^2 (n_2^2-1)^{1/4} (r^2 n^2 - a^2)^{1/4} (r_0^2 n^2 - a^2)^{1/4}} \cdot$$

$$\exp\{-ST_s - x_j (sa/\alpha)^{1/3}\theta_s\}, \qquad (r \geqslant a) \tag{34}$$

$$\bar{\phi}_j(r,\theta,s) \sim -6\sum_{j=1}^{\infty} \frac{\pi^{1/2}\alpha^{7/6} a^{1/6}}{s^{5/6}} \frac{(n^2-1)^{1/4}}{f'(x_j)(n^2-2)(r^2-a^2)^{1/4}(r_0^2 n^2 - a^2)^{1/4}}$$

$$\exp\{-ST_p - x_j (sa/\alpha)^{1/3}\theta p\}, \qquad (r \geqslant a) \tag{35}$$

$$\psi_j^{(2)}(r,\theta,s) \sim -\frac{\sqrt{3}}{2}\sum_{j=1}^{\infty} \pi a^{1/3}\left(\frac{\beta}{s}\right)^{2/3} \frac{1}{(r^2-a^2)^{1/4}(r_0^2-a^2)^{1/4}} \cdot$$

$$\frac{F(x_j)}{f'(x_j)} \exp\{-ST - x_j (sa/\beta)^{1/3}\delta\}, \qquad (r \geqslant a) \tag{36}$$

where $$T_s = \frac{1}{\beta}\left\{\left(r_0^2 - \frac{a^2}{n^2}\right)^{1/2} - 2a\left(1 - \frac{1}{n^2}\right)^{1/2} + \left(r^2 - \frac{a^2}{n^2}\right)^{1/2}\right\} + \frac{a\theta_s}{\alpha}, \tag{37}$$

$$T_p = \frac{1}{\beta}\left\{\left(r_0^2 - \frac{a^2}{n^2}\right)^{1/2} - a\left(1 - \frac{1}{n^2}\right)^{1/2}\right\} + \frac{1}{\alpha}\{(r^2-a^2)^{1/2} + a\theta p\}, \tag{38}$$

$$T = \frac{1}{\beta}\{(r_0^2 - a^2)^{1/2} + (r^2-a^2)^{1/2} + a\delta\}, \tag{39}$$

and $$\theta_s = |\theta| - \cos^{-1}(a/nr) - \cos^{-1}(a/nr_0) + 2\cos^{-1}(1/n), \tag{40}$$

$$\theta_p = |\theta| - \cos^{-1}(a/r) - \cos^{-1}(a/nr_0) + \cos^{-1}(1/n), \tag{41}$$

$$\delta = |\theta| - \cos^{-1}(a/r) - \cos^{-1}(a/r_0). \tag{42}$$

We will now perform Laplace inversion to determine the pulse propagation modes. For this, first of all, we shall substitute the approximations of the Laplace transform of the mode solution in the series (21) and (22) and perform term-by-term inversion. We know that term-by-term inversion is quite reasonable in a certain physically important region called ' deep shadow ' [3]. To determine this region, let us consider [34]. Each $\bar{\psi}_j^{(1)}$ contains a factor exp $(-sT_s)$. This may be called a delay factor because $\psi_j^{(1)}$, if it exists, is zero for $t < T_s$ and it can be obtained for $t > T_s$ by replacing t by $t - T_s$ in the inverse Laplace transform of exp $(sT_s)\,\bar{\psi}_j^{(1)}$ according to shift-rule. However, the application of the shift-rule presupposes that exp $(ST_s)\,\bar{\psi}_j^{(1)}$, is a Laplace transform and consequently it must tend to zero as $s \to \infty$. This is possible, only if $\theta_s > 0$. The same argument applies to (35) and (36). Therefore (34) can be used only in the region

$$\theta_s \geq 0, \tag{43}$$

and (35) and (36) can be used only in the region

$$\theta_p \geq 0, \tag{44}$$

and $\delta \geq 0$, respectively.

Friedlander [4] has shown that the leading term of any mode solution of order $j > 1$ tends to zero more rapidly than the error term of the first mode. The arguments are exactly the same in the present case. Therefore it seems quite reasonable to retain only the leading term of the first mode when we use our approximations. Now we can obtain the Laplace inversion of $\bar{\psi}_j$ and $\bar{\phi}_j$ explicitly if we use the following formula

$$\frac{1}{2\pi i}\int_{\sigma-i\infty}^{\sigma+i\infty} \bar{s}^k \exp\{ST - \gamma s^{1/3}\}\,ds \sim \frac{3^{(6k-1)/4}}{2\pi^{1/2}}\,\gamma^{(3-6k)/4}\,T^{(6k-5)/4}\,H(T).\exp\left(\frac{-2\gamma^{3/2}}{3^{3/2}\,T^{1/2}}\right). \tag{45}$$

$H(T)$ is the Heaviside unit function and K as well as γ are real and positive. Using (45), we get from (34), (35) and (36)

$$\psi^{(1)}(r,\theta,t) \sim -\frac{3^{5/4}\pi^{1/2}\alpha^{19/12}}{2a^{1/4}(n^2-2)^2}\left[\frac{4\rho\beta^4(n^2-1)^{1/2}(n_2^2-1)^{1/2} - \rho'\alpha^4}{\rho\beta^4(n_2^2-1)^{1/4}}\right] \times \frac{T_1^{1/4}H(T_1)}{\{(n_2^2-1)(r^2n^2-a^2)(r_0^2n^2-a^2)\}^{1/4}(x_1\theta s)^{3/4}} \times \exp\left\{-2\left(\frac{x_1\theta s}{3}\right)^{3/2}\left(\frac{a}{\alpha T_1}\right)^{1/2}\right\},\ (r \geq a) \tag{46}$$

$$\phi(r,\theta,t) \sim \left[-\frac{9a^{4/3}(n^2-1)^{1/4}}{(n^2-2)(r^2-a^2)^{1/4}(r_0^2n^2-a^2)^{1/4}}\cdot \frac{1}{f'(x_1)(x_1\theta p)^{1/2}}\right] \times H(T_2)\exp\left\{-2\left(\frac{x_1\theta p}{3}\right)^{3/2}\left(\frac{a}{aT_2}\right)^{1/2}\right\}, \quad (r \geqslant a) \tag{47}$$

$$\psi^2(r,\theta,t) \sim \left[-\frac{3^{5/4}\pi^{1/2}a^{1/4}\beta^{3/4}}{4\{(r^2-a^2)(r_0^2-a^2)x_1\delta T_3\}^{1/4}}\frac{F(x_1)}{f'(x_1)}\right] \times H(T_3)\exp\left\{-2\left(\frac{x_1\delta}{3}\right)^{3/2}\left(\frac{a}{\beta T_3}\right)^{1/2}\right\}, \quad (r \geqslant a) \tag{48}$$

where
$$T_1 = t - T_s,$$
$$T_2 = t - T_p, \tag{49}$$
$$T_3 = t - T_\bullet.$$

To interpret physically the solutions obtained in (46), (47) and (48), we first give a brief description of the geometry of the problem. Initially the incident SV pulse striking the outer surface of the cylinder gives rise to reflected S, reflected P and refracted P pulses according to the laws of ordinary geometrical optics (figure 2). When the rays strike the outer surface of the cylinder at the critical angle, the

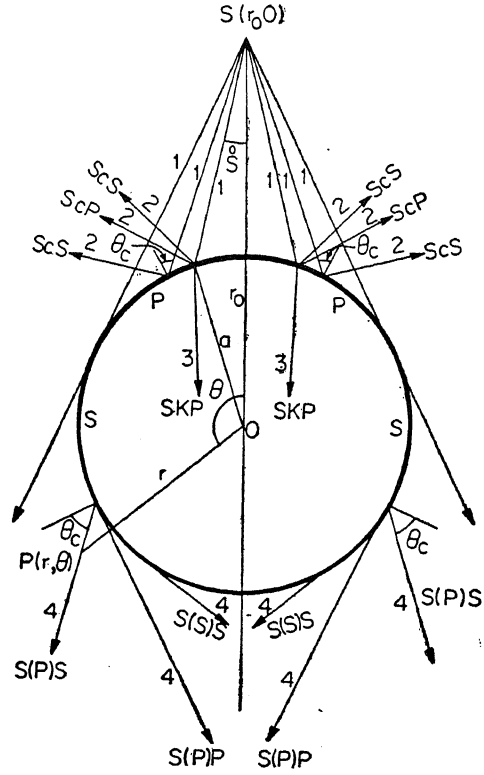

Figure 2. 1. Incident ray, 2. Reflected ray, 3. Refracted ray, and 4. Diffracted ray.

reflected P rays become tangential to the surface and as such they move along the surface. These surface waves at each point of their path give rise to lateral SV waves at critical angle in the outer medium and P waves along the tangent to the surface [8]. The former are denoted by $S(P)S$ and the latter by $S(P)P^6$. In the case of grazing incidence, the disturbances move along the surface and at each point of their path they shed diffracted SV waves in the outer medium tangential to the surface. These are denoted by $S(S)S$ [1].

Now we consider the approximate solution we have obtained. If we interpret them in the light of Keller's geometrical theory of diffraction [8], we find that [46] represents the $S(P)S$ diffraction event while (47) represents the $S(P)P$ diffraction event. The expression given by (48) stands for the $S(S)S$ diffracted event.

6. Conclusions

We have obtained the solution in terms of pulse propagation modes outside the cylinder and interpreted them as diffracted pulses. The arrival times of the diffracted pulses agree with those predicted by Keller's geometrical theory of diffraction. We note that the arrival time of the diffracted pulses is the same whether the obstacle is a rigid, weak or fluid cylinder.

It has been conjectured by Friedlander [3] that near the diffracted front, the diffracted pulse due to a continuously curved obstacle is always of the form $AT^k \exp(-\sigma T^{-\mu})$, where T is the time counted from the arrival of the diffracted front, A and σ are functions of position and K and μ are constants such that $-3/2 \leqslant K \leqslant 1$ and $0 < \mu \leqslant 1$. We find that this is true in our case.

In the illuminated region, the reflected and refracted pulses can also be obtained by using Friedlander's method. We have left this case here to avoid a lengthy paper as the calculation of these pulses becomes rather involved. We hope to investigate this case separately.

References

[1] Bullen K E 1963 *Theory of seismology* (Cambridge : University Press)
[2] Friedlander F G 1954 *Commun. Pure Appl. Math.* **7** 705
[3] Friedlander F G 1955 Electromag. Res. Inst. Math. Sci., New York, Res. Rep. EM–76, MR 16, 977.
[4] Friedlander F G 1958 *Sound pulses* (Cambridge : University Press)
[5] Gilbert F 1960 *J. Acoust. Soc. Am.* **32** 841
[6] Gilbert F and Knopoff L 1959 *J. Acoust. Soc. Am.* **31** 1169
[7] Jha R 1974 *Proc. Cambridge Philos. Soc.* **75** 391
[8] Keller J B 1958 Div. Electromag. Res. Inst. Math. Sci., New York Univ. Res. Rep. EM–115
[9] Mishra S K 1964a *Proc. Cambridge Philos. Soc.* **60** 295
[10] Mishra S K 1964b *Proc. Indian Acad. Sci.* **A59** 21
[11] Olver F W J 1954 *Philos. Trans. R. Soc. (London)* **A247** 328
[12] Watson G N 1944 *Theory of Bessel functions* (Cambridge : University Press)
[13] White J E 1965 *Seismic waves : Radiation, transmission and attenuation* (New York : McGraw-Hill)

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 3, September 1980, pp. 171–181.

Numerical solutions of the improved Boussinesq equation

LABIB ISKANDAR and PADAM C JAIN
Department of Mathematics, Indian Institute of Technology, Bombay 400 076, India

MS received 8 March 1980

Abstract. The combined approach of linearisation and finite difference method is used to solve the improved Boussinesq equation. A three-level iterative scheme having second order accuracy and constant coefficients matrix is devised and used in discussing the dynamics of waves having various initial wave packets. The results are in good agreement with the available results.

Keywords. Nonlinear equation; Boussinesq equation; numerical solution.

1. Introduction

The Boussinesq equation

$$u_{tt} = u_{xx} + (u^2)_{xx} + u_{xxxx}, \tag{1}$$

describes a large number of nonlinear dispersive wave phenomena, such as (i) propagation of long waves on the surface of shallow water in both directions, (ii) propagation of long waves in one-dimensional nonlinear lattices, (iii) propagation of ion-sound(s) waves in a uniform isotropic plasma.

Bogolubsky [3, 4] has shown that equation (1) describes an unphysical instability of short wavelengths, and the Cauchy problem for this equation is incorrect. Keeping the dynamics of the long wavelengths unchanged, Bogolubsky has proposed the Improved Boussinesq (IBq.) equation

$$u_{tt} = u_{xx} + (u^2)_{xx} + u_{xxtt}, \tag{2}$$

which is physically stable, correct in the sense of Cauchy problem and convenient for computer simulation. Bogolubsky has used ordinary implicit finite difference scheme to study the dynamics of various initial wave packets. He has demonstrated that solitary waves interaction described by (2) is inelastic and that the coefficient of inelasticity increases with an increase of amplitude of the interacting solitary waves.

In this paper, a combined approach [6, 7] of linearisation technique and finite difference method is presented for computing solutions of the improved Boussinesq equation (2).

2. Finite difference scheme

Let $u^{(n)}$ denote the value of u at the nth iteration and $u^{(0)}$ be the initial guess. We consider the following approximation for equation (2) :

$$u_{tt}^{(n+1)} = u_{xx}^{(n+1)} + u_{xxtt}^{(n+1)} + (u^2)_{xx}^{(n)}, n = 0, 1, 2, \ldots . \tag{3}$$

The functions in the sequence $\{u^{(n)}\}$ satisfy the boundary conditions specified for u. The linear convergence of the sequence $\{u^{(n)}\}$ to the solution of the original nonlinear problem has been established by Bellman and Kalaba [2]. The sequence of the linear problems given by equation (3) along with the given initial and boundary conditions is solved by using the finite difference method.

For the computational work, we consider a finite interval on the x-axis. The domain in $x - t$ plane is discretised by a grid with step length h and time step k. The exact value of u at the grid point $(x, t) \equiv (rh, sk)$ is denoted by $u_{r,s}$ ($r = 0, 1, 2, \ldots, N$ and $s = 0, 1, 2, \ldots$) while the numerical value is designated by $U_{r,s}$. We take

$$u_{tt}|_{r,s} = \frac{1}{k^2}(u_{r,s+1} - 2u_{r,s} + u_{r,s-1}) + 0(k^2), \tag{4}$$

$$u_{xx}|_{r,s} = \frac{1}{2h^2}(u_{r+1,s+1} - 2u_{r,s+1} + u_{r-1,s+1} + u_{r+1,s-1} - 2u_{r,s-1} + u_{r-1,s-1}) + 0(h^2 + k^2), \tag{5}$$

$$u_{xxtt}|_{r,s} = \frac{1}{h^2 k^2}[(\delta^2 u_r)_{s+1} - 2(\delta^2 u_r)_s + (\delta^2 u_r)_{s-1}] + 0(h^2 + k^2), \tag{6}$$

where $(\delta^2 u_r)_s = u_{r+1,s} - 2u_{r,s} + u_{r-1,s}$.

$$(u^2)_{xx}|_{r,s} = \frac{1}{2h^2}[(\delta^2 u_r^2)_{s+1} + (\delta^2 u_r^2)_{s-1}] + 0(h^2 + k^2) \tag{7}$$

where $(\delta^2 u_r^2)_s = (u_{r+1,s})^2 - 2(u_{r,s})^2 + (u_{r-1,s})^2$.

Substituting equations (4) to (7) in equation (3) and neglecting the truncation error, the resulting algebraic system of equations takes the form

$$\begin{aligned} &- D(U_{r+1,s+1}^{(n+1)} + U_{r-1,s+1}^{(n+1)}) + HU_{r,s+1}^{(n+1)} \\ &= P[(U_{r+1,s+1}^{(n)})^2 - 2(U_{r,s+1}^{(n)})^2 + (U_{r-1,s+1}^{(n)})^2] \\ &\quad - \frac{2}{h^2}(U_{r+1,s}^{(n_s)} + U_{r-1,s}^{(n_s)}) + 2\left(1 + \frac{2}{h^2}\right)U_{r,s}^{(n_s)} \\ &\quad + P[(U_{r+1,s-1}^{(n_{s-1})})^2 - 2(U_{r,s-1}^{(n_{s-1})})^2 + (U_{r-1,s-1}^{(n_{s-1})})^2] \\ &\quad + D(U_{r+1,s-1}^{(n_{s-1})} + U_{r-1,s-1}^{(n_{s-1})}) - HU_{r,s-1}^{(n_{s-1})}, \\ &r = 1, 2, \ldots, N-1, \quad s = 1, 2, \ldots, \quad n = 0, 1, 2, \ldots, \end{aligned} \tag{8}$$

where $$D = \frac{k^2}{2h^2} + \frac{1}{h^2}, \quad H = 1 + \frac{k^2}{h^2} + \frac{2}{h^2}, \quad P = \frac{k^2}{2h^2}, \tag{9}$$

and the superscript n_s denotes the final number of iterations required to obtain an acceptable approximation to the value of $U_{r,s}$ at the grid points on the line $t = sk$ subject to the criterion [2]

$$\max_r \mid U_{r,s}^{(n+1)} - U_{r,s}^{(n)} \mid \leq 10^{-5}, \quad 1 \leq r \leq N. \tag{10}$$

A similar definition is valid for the superscript n_{s-1}. To investigate the stability of scheme (8), we apply Fourier stability method [1]. We drop iteration index and linearise (8) by taking $(U_{r,s+1})^2$, $(U_{r,s})^2$ and $(U_{r,s-1})^2$ as $MU_{r,s+1}$, $MU_{r,s}$ and $MU_{r,s-1}$ with M as a constant such that

$$\max_r \{ \mid U_{r,s+1} \mid, \quad \mid U_{r,s} \mid, \quad \mid U_{r,s-1} \mid \} < \mid M \mid.$$

Substituting

$$U_{r,s} = \psi^s e^{i(mrh)}, \quad i = \sqrt{-1} \text{ in (8),}$$

we get

$$\psi^2 - 2\rho\psi + 1 = 0, \tag{11}$$

where $$\rho = \left[1 + \frac{4\sin^2\phi}{h^2}\right] \Big/ \left[1 + \frac{4\sin^2\phi}{h^2} + 2(1+M)\frac{k^2\sin^2\phi}{h^2}\right] \tag{12}$$

and $\phi = mh/2$.

The stability condition is

$$\left| \frac{1 + \dfrac{4\sin^2\phi}{h^2}}{1 + \dfrac{4\sin^2\phi}{h^2} + 2(1+M)\dfrac{k^2\sin^2\phi}{h^2}} \right| \leq 1, \text{ for all } \phi. \tag{13}$$

It is satisfied for arbitrary values of k/h. Hence, the linearised form of the difference scheme (8) is unconditionally stable. The local truncation error of scheme (8) is found to be $0(k^2 + h^2)$.

3. Numerical methods

Consider the initial conditions

$$u_{r,0} = f(rh), \quad u_t|_{r,0} = g(rh), \tag{14}$$

and the boundary values $u_{0,s}$ and $u_{N,s}$ are specified for all s. The first initial condition specifies $U_{r,0}$ on the line $t = 0$. We use the second initial condition to find values on the line $t = k$ by employing a false boundary and the second-order central difference formula

$$u_t|_{r,0} = \frac{u_{r,1} - u_{r,-1}}{2k} + 0(k^2). \tag{15}$$

Writing $g(rh) = g_r$, we have the approximation

$$U_{r,1} - U_{r,-1} = 2k\, g_r. \tag{16}$$

For simplicity, we rewrite scheme (8) in the following form

$$-D(W_{r+1}+W_{r-1})+HW_r=F,$$

$$r=1,\ 2,\ \cdots,N-1,\ s=1,\ 2,\ \cdots,\ n=0,\ 1,\ 2,\ \cdots, \tag{17}$$

where $$F=P(Y_{r+1}^2-2Y_r^2+Y_{r-1}^2)-\frac{2}{h^2}(U_{r+1}+U_{r-1})+2\left(1+\frac{2}{h^2}\right)U_r$$

$$+P(V_{r+1}^2-2V_r^2+V_{r-1}^2)+D(V_{r+1}+V_{r-1})-HV_r, \tag{18}$$

D, H, P are defined in (9) and

$$V_r\equiv U_{r,\,s-1}^{(n+\frac{1}{2})},\quad U_r\equiv U_{r,\,s}^{(n+\frac{1}{2})},\quad W_r\equiv U_{r,\,s+1}^{(n+1)},\quad Y_r\equiv U_{r,\,s+1}^{(n)}. \tag{19}$$

From (16) replacing $U_{r,-1}$ $(\equiv V_r)$ by (W_r-2kg_r) in the linear terms of (17), and by (Y_r-2kg_r) in the nonlinear terms of (17), we get

$$-D(W_{r+1}+W_{r-1})+HW_r=G,$$

$$r=1,\ 2,\ \cdots,N-1;\ s=1;\ n=0,\ 1,\ 2,\ \cdots, \tag{20}$$

where $$G=-\frac{1}{h^2}(U_{r+1}+U_{r-1})+\left(1+\frac{2}{h^2}\right)U_r+P(Y_{r+1}^2-2Y_r^2+Y_{r-1}^2)$$

$$+kHg_r-KD(g_{r+1}+g_{r-1})+2k^2P(g_{r+1}^2-2g_r^2+g_{r-1}^2)$$

$$-2kP(g_{r+1}Y_{r+1}-2g_rY_r+g_{r-1}Y_{r-1}). \tag{21}$$

We solve the tridiagonal system of linear equations (20) subject to the known initial values $U_{r,0}$, for all r and the known boundary values $U_{0,1}$ and $U_{N,1}$, by Thomas algorithm [1], thereby eliminating the matrix operations.

At this first-time level, we follow an iterative procedure : (i) Put $Y_r=U_r$, for all r; (ii) solve (20) for W_r as mentioned above; (iii) take $Y_r=W_r$, for all r; (iv) calculate improved values of W_r from (20); (v) repeat the iterative procedure (i) to (iv) till the condition (10) is satisfied.

It may be noted that iterative scheme (20) is a two-level scheme having second-order accuracy and has a constant coefficient matrix.

For the time levels $\{s+1 \mid s=1,\ 2,\ 3,\ \cdots\}$, we solve the three-level iterative scheme (17) subject to the known initial values $U_{r,s}$ and $U_{r,s-1}$ for all r and the known boundary values $U_{0,s}$ and $U_{N,s}$ for all s. Scheme (17) can be solved for W_r in the same way as scheme (20).

The three-level iterative scheme (17) has the advantage of having second order accuracy and requires the solution of a constant coefficient tridiagonal system at each time level.

4. Numerical examples and results

In this section, we solve four test examples to illustrate the efficiency of the proposed numerical method in studying the nonlinear dynamics of the solution of the modified Boussinesq equation.

Example 1

Equation (2) has a special solution [3] which describes solitary waves (soliton type solutions) :

$$u(x, t) = A \operatorname{sech}^2\left[\frac{1}{M}\left(\frac{A}{6}\right)^{1/2}(x - Mt - x_0)\right], \tag{22}$$

where A, x_0 are arbitrary constants and

$$M = \left(1 + \frac{2}{3}A\right)^{1/2}. \tag{23}$$

Consider equation (2) with the following initial data :

$$u(x, 0) = A \operatorname{sech}^2\left[\frac{1}{M}\left(\frac{A}{6}\right)^{1/2} x\right], \tag{24}$$

and $$u_t(x, 0) = 0. \tag{25}$$

The boundaries on the right and left hand sides have been taken to fit $u \approx 0$. For the computational work, we have chosen $A = 0{\cdot}5$, $k = h = 0{\cdot}5$ and $-250 \leq x \leq 250$. Numerical computations were carried out upto $t = 200$. The results show a symmetric pattern of the wave propagation. Figure 1 represents half the picture of this propagation of the wave. The initial stationary packet given by (24) and (25) is disintegrated into two symmetric solitary waves of smaller amplitudes moving in opposite directions with oscillating tail between the diverging solitary waves. The amplitude of the leading solitary wave decreases with time and the tail spreads over a large distance. The number of iterations required to satisfy condition (10) at each time level was three. By taking the value of M greater than the value given by (23) for a fixed A, it is likely that the number of solitary waves formed decreases. But in our computational work, we got two solitary waves in conformity with the result obtained by Bogolubsky [3].

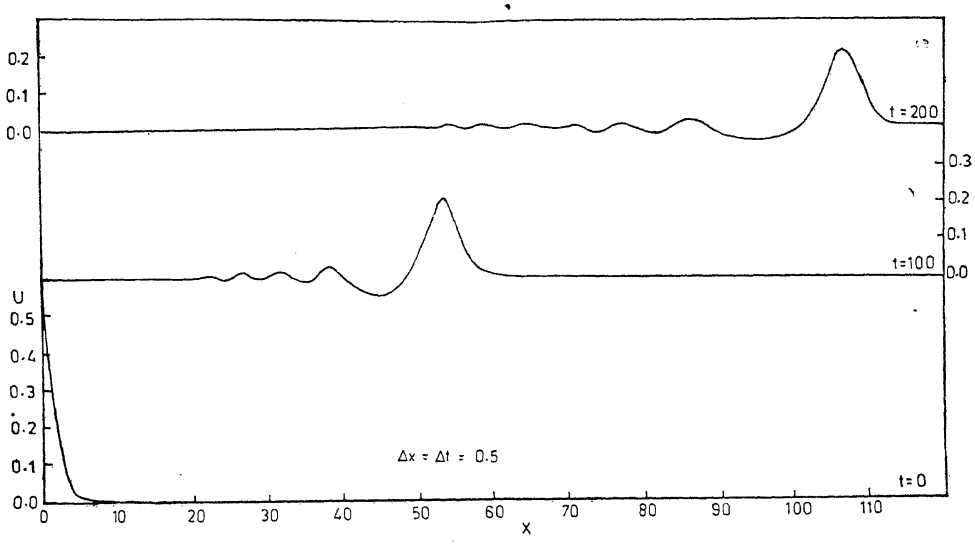

Figure 1. IBq. solitary waves formed from symmetric perturbation at rest (half picture is shown),

Example 2

The interaction of two IBq. solitary waves having different amplitudes and travelling in the same direction has been discussed in this example. We start with initial conditions given by the linear sum of two well-separated solitary waves of different amplitudes :

$$u(x, 0) = A_1 \operatorname{sech}^2\left[\frac{1}{M_1}(A_1/6)^{1/2} x + c\right] + A_2 \operatorname{sech}^2\left[\frac{1}{M_2}(A_2/6)^{1/2} x - c\right] \tag{26}$$

and

$$u_t(x, 0) = 2A_1 (A_1/6)^{1/2} \operatorname{sech}^2\left[\frac{1}{M_1}(A_1/6)^{1/2} x + c\right] \times \tanh\left[\frac{1}{M_1}(A_1/6)^{1/2} x + c\right] + 2A_2 (A_2/6)^{1/2} \operatorname{sech}^2\left[\frac{1}{M_2}(A_2/6)^{1/2} x - c\right] \times \tanh\left[\frac{1}{M_2}(A_2/6)^{1/2} x - c\right], \tag{27}$$

where $M_1 = \left(1 + \frac{2}{3} A_1\right)^{1/2}$, $M_2 = \left(1 + \frac{2}{3} A_2\right)^{1/2}$

In order to make the range of x finite, we cut the tails of the hyperbolic secant (22). After some numerical experiments, it was found that $c = 4{\cdot}0$ was a good value for maintaining the accuracy within reasonable time of computation. The boundaries on the right and left hand sides have been taken to fit $u \approx 0$.

We have taken $A_1 = 2{\cdot}0$, $A_2 = 0{\cdot}5$, $k = h = 0{\cdot}5$ and $0 \leq x \leq 250$. A large range of x has been considered to get the true behaviour of the interaction between the two solitary waves. The number of iterations taken for the solution to converge at each time level was seven. The two solitary wave interaction, plotted at different time steps up to the time level 275, is shown in figure 2. It is seen that in the nonlinear interaction region, the maximum joint amplitude is less than that of the largest solitary wave and there is a phase shift in each solitary wave resulting from the interaction. We may conclude that the solitary wave interaction for the improved Boussinesq equation is not clean due to the development of the oscillatory tail which spreads over a large distance with time.

Example 3

Now we study the interaction of two solitary waves of equal amplitudes and moving in opposite directions (head-on collision). We discuss two cases when the solitary wave amplitude $A \ll 1$ and $A \geq 1$.

Case I. Using the dispersion relation of the linearised form of equations (1) and (2), Bogolubsky [4] has shown that these two equations yield similar results for $A \ll 1$. Hence, we expect that the interaction between two solitary waves with small amplitude ($A \ll 1$) would differ only slightly when described by equations (1) and (2).

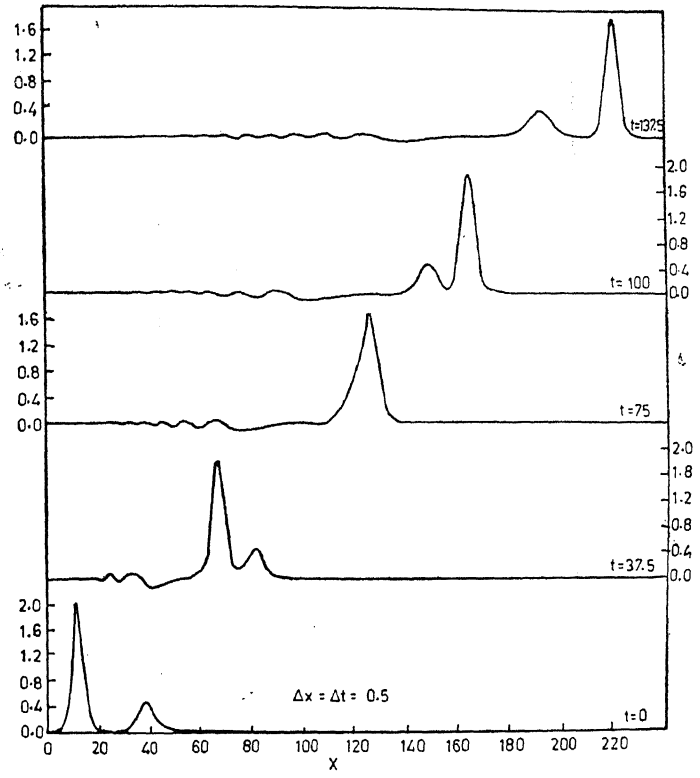

Figure 2. IBq. solitary wave interaction for $A_1 = 2{\cdot}0$, $A_2 = 0{\cdot}5$ at times $t = 0$, $37{\cdot}5$, 75, 100, $137{\cdot}5$.

We consider the initial conditions (26) and

$$u_t(x, 0) = 2A_1\left(\frac{A_1}{6}\right)^{1/2} \operatorname{sech}^2\left[\frac{1}{M_1}(A_1/6)^{1/2}x + c\right] \times \tanh\left[\frac{1}{M_1}(A_1/6)^{1/2}x + c\right] - 2A_2\left(\frac{A_2}{6}\right)^{1/2} \operatorname{sech}^2\left[\frac{1}{M_2}(A_2/6)^{1/2}x - c\right] \times \tanh\left[\frac{1}{M_2}(A_2/6)^{1/2}x - c\right]. \tag{28}$$

The boundaries on the right and left hand sides have been taken to fit $u \approx 0$.

We have taken $A_1 = A_2 = 0{\cdot}1$, $k = 0{\cdot}25$ and $h = 0{\cdot}5$. After some numerical trials, it was found that $c = 4{\cdot}0$ and $-65{\cdot}5 \leq x \leq 65{\cdot}5$ were reasonable values for computational work. In the present case, two iterations were required for the solution to converge at each time level. In figure 3, the head-on collision is plotted at different time steps.

When the two solitary waves overlap, then the joint amplitude is smaller than twice the amplitude of an individual solitary wave. The interaction is clean and exhibits true soliton behaviour (elastic interaction).

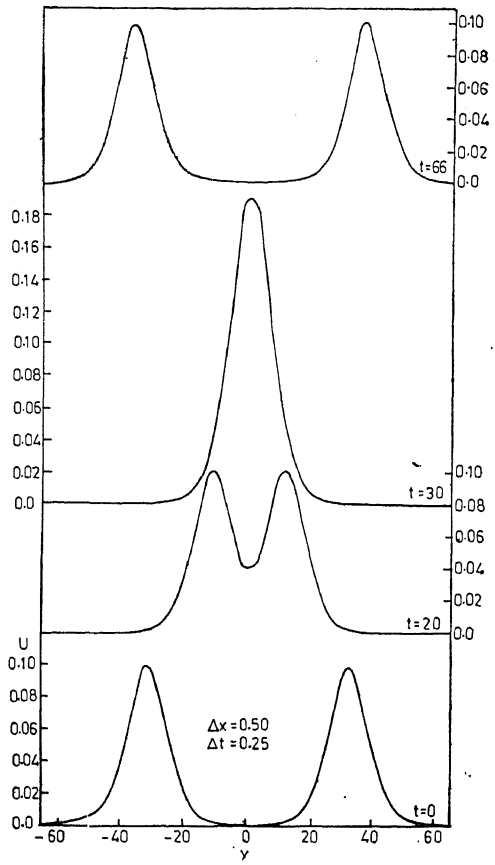

Figure 3. Head-on collision of IBq. solitary waves for $A_1 = A_2 = 0{\cdot}1$.

Similar results have been obtained analytically by Hirota [5] for the Boussinesq equation (1). This confirms the conclusion that equations (1) and (2) yield similar results for $A \ll 1$.

Case II. In this case, we consider large values of the amplitude A up to $A = 2$. The initial and boundary conditions are chosen to be the same as those prescribed in Case I. For a representative value of $A = 1{\cdot}0$, figures 4 (a) and 4 (b) show the head-on collision of these identical solitary waves. In all the experiments with different values of A, the numerical results obtained from the solution of (2) with identical solitary waves moving towards each other, the joint amplitude of these solitary waves was found to be smaller than twice the amplitude of an individual solitary wave. It was observed that the amplitude of the solitary wave does not change appreciably as a result of the interaction. In between the region of two diverging solitary waves, one finds the solution which alternates in sign and consists of almost linear waves whose amplitude is small compared with that of the solitary waves and whose wavelength is of the order of the initial width of the solitary wave.

Bogolubsky [4] defines the inelasticity coefficient K_{in} as ' The ratio of the amplitude of the non-soliton perturbations produced in soliton collisions to their

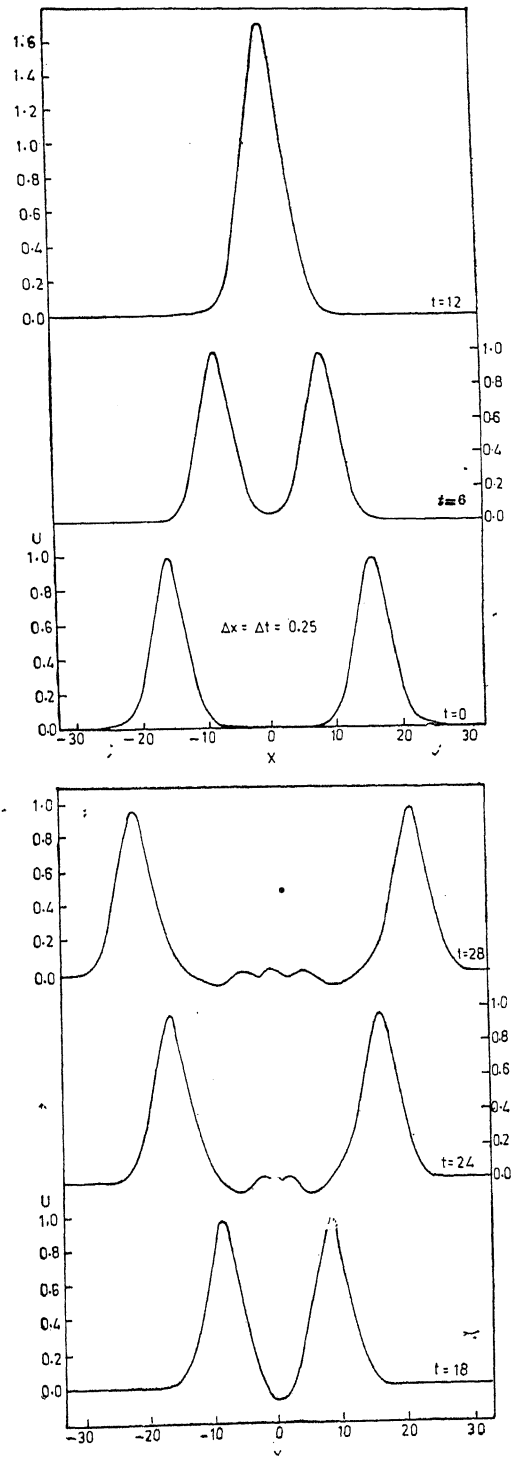

Figure 4. Head-on collision of IBq. solitary waves for $A_1 = A_2 = 1{\cdot}0$, at times. **a.** $t = 0, 6, 12$. **b.** $t = 18, 24, 28$.

amplitude A'. Figure 5 shows that the coefficient of inelasticity increases with an increase in the amplitude of the interacting solitary waves.

Example 4

In this example, we discuss the dynamics of the interaction of solitary waves having different amplitudes and colliding head-on. The initial and boundary conditions are taken to be the same as for example 3. For numerical computations, we have taken $k = h = 0{\cdot}25$, $c = 5{\cdot}0$ and $-32{\cdot}75 \leq x \leq 32{\cdot}75$. A representative case is plotted in figure 6 for $A_1 = 1{\cdot}0$, $A_2 = 2{\cdot}0$. In this case, four iterations were needed at each time level for the solution to converge. The perturbation which appears in the region between two diverging solitary waves increases as the

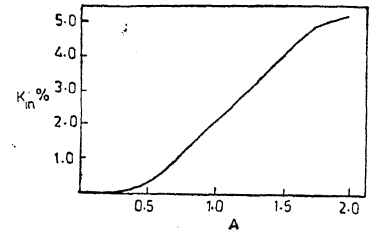

Figure 5. Dependence of K_{4n} on A for the IBq. equation.

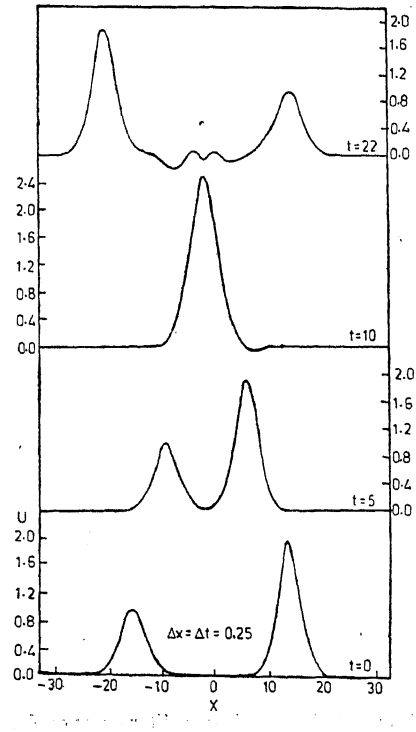

Figure 6. Head-on collision of IBq. solitary waves for $A_1 = 1{\cdot}0$ and $A_2 = 2{\cdot}0$.

amplitude of the interacting solitary waves increases, while the amplitude of the individual solitary wave does not change appreciably due to interaction.

5. Conclusions

Our study of the solitary wave interaction of IBq. equation shows that :

(i) The solitary waves with amplitudes below a critical value interact with each other elastically.

(ii) The solitary waves with amplitudes above the critical value interact with each other inelastically.

(iii) The coefficient of inelasticity increases with the increase in the amplitude of the interacting solitary waves.

In the literature, one finds a few papers dealing with the solution of IBq. equation, but without the details of computational results. In the present paper, we have discussed four examples to get an insight into the properties of the solutions of IBq. equation. The numerical method developed by us is iterative while the others are not; our method provides results which are in good conformity with the results obtained by Bogolubsky [3].

References

[1] Ames W F 1977 *Numerical methods for partial differential equations* II ed. (New York : Academic Press)

[2] Bellman R E and Kalaba R E 1965 *Quasilinearisation and nonlinear boundary value problems* (New York : American Elsevier)

[3] Bogolubsky I L 1977 *Comput. Phys. Commun.* **13** 149

[4] Bogolubsky I L 1976 *JETP Lett.* **24** 160

[5] Hirota R 1973 *J. Math. Phys.* **14** 810

[6] Iskandar L and Jain P C 1980 *Comp. and Maths. with Appls.* (in the press)

[7] Jain P C and Iskandar L 1979 *Comp. Maths. Appl. Mech. Eng.* **20** 195

Proc. Indian Acad. Sci. (Math. Sci.), Vol. 89, Number 3, September 1980, pp. 183–192.

Propagation of discontinuities along bicharacteristics in the unsteady flow of a relaxing gas

V D SHARMA* and RADHE SHYAM

Applied Mathematics Section, Institute of Technology, Banaras Hindu University, Varanasi 221 005, India

*Present address: Department of Aerospace Engineering, University of Maryland, College Park, Maryland 20742, USA

MS received 22 December 1978 ; revised 5 May 1980

Abstract. Growth and decay of weak discontinuities headed by wave front of arbitrary shape in three dimensions are investigated in an unsteady flow of a relaxing gas. The transport equations representing the rate of change of discontinuities in the normal derivatives of the flow variables are obtained and it is found that the nonlinearity in the governing equations plays an important role in the interplay of damping and steepening tendencies of the wave. An explicit criterion for the growth and decay of weak discontinuities along bicharacteristic curves in the characteristic manifold of the governing differential equations is given and special reference is made of diverging and converging waves under different thermodynamical situations. It is shown that there is a special case of a compressive converging wave, irrespective of the thermodynamical state whether weak or strong, in which the effects of thermodynamical influences and that of wave front curvature are unable to overcome the tendency of the wave to grow into a shock.

Keywords. Strong equilibrium; weak equilibrium; weak discontinuity; weak wave; bicharacteristic curve; diverging waves; converging waves.

1. Introduction

Bowen [1] discussed the propagation of plane acceleration waves in a mixture of chemically reacting elastic materials. Calling a state with a zero reaction rate and a non-zero affinity a "weak equilibrium state", and one with both these quantities zero a ' strong equilibrium state", Bowen showed that strong equilibrium states have a stabilising effect in that not all compressive waves can grow into shock waves. Bowen also showed that weak equilibrium states have the property that plane compressive waves, no matter how small initially, can grow without bound. Bowen considered only plane waves and did not determine the effects of curvature on the propagation of acceleration waves. The propagation of weak discontinuities through strong equilibrium states (in the terminology of Bowen [1]) in one-dimensional flow of a relaxing gas has been discussed by a number of investigators (a short account is given by Clarke and McChesney [3], § 2·4). But waves of arbitrary shape have not so far been studied. Recently, Elcrat [5] has

used the combination of singular surface theory and the theory of bicharacteristics to study the growth of sonic discontinuities in an unsteady flow of a perfect gas. An application of this method is made in this paper to study the growth and decay behaviour of weak discontinuities of arbitrary shape along the bicharacteristic curves in the characteristic manifold of the differential equations governing an unsteady flow of a relaxing gas. The situation in which the medium ahead of the wave is uniform and at rest has been envisaged and special reference is made of diverging and converging waves. The effects of thermodynamical influences and that of wave front curvature on the growth and decay properties of the diverging and converging waves are studied. It is found that, in a strong equilibrium state, all compressive waves, except in one special case of converging waves in which both the initial principal curvatures are positive and equal, grow without bound only if the magnitude of the initial discontinuity associated with the wave exceeds a critical value. In this special case, it is shown that a compressive wave, no matter how weak initially, always grows into a shock before the formation of the focus, irrespective of the thermodynamical state whether weak or strong. Further, in a weak equilibrium state all diverging compressive waves, no matter how weak initially, always end up into a shock after a finite time.

2. Basic equations

In the absence of viscosity, heat conduction and body forces, the equations governing an unsteady flow of a relaxing gas are ([2], p. 37)

$$\frac{\partial p}{\partial t} + u_i\, p_{,i} + \rho c_f^2\, u_{i,i} = -\sigma\omega\, c_f^2\,, \tag{1}$$

$$\rho\,\frac{\partial u_i}{\partial t} + \rho\, u_j\, u_{i,j} + p_{,i} = 0, \tag{2}$$

$$\frac{\partial S}{\partial t} + u_i\, S_{,i} = \left(\frac{-\,\partial h/\partial q}{\partial h/\partial S}\right)\omega\,, \tag{3}$$

$$\frac{\partial q}{\partial t} + u_i\, q_{,i} = \omega\,, \tag{4}$$

where the summation convention on repeated indices is employed, and a comma followed by an index (say i) denotes a partial derivative with respect to a space variable x_i. The range of Latin indices is taken to be 1, 2, 3. The symbols appearing in (1) to (4) are as follows : u_i are the gas velocity components, p is the pressure, ρ is the density, $h = h(p, S, q)$ is the enthalpy of the gas, S is the entropy per unit mass of the medium, q is the progress variable characterising the extent of internal transformation in the fluid, $(-\,\partial h/\partial q)$ is the chemical affinity of the internal transformation, $\omega = \omega(p, S, q)$ is the rate of internal transformation, $c_f = (\partial p/\partial \rho)_{S,q}^{1/2}$ is the frozen sound speed, t is the time and

$$\sigma = \left(\frac{-\,\partial h/\partial q}{\partial h/\partial S}\right)\frac{\partial \rho}{\partial S} + \frac{\partial p}{\partial q}\,.$$

Equations (1) to (4), which form a set of six equations in six unknowns (u_i, p, S, q), are sufficient to discuss the problem under consideration. Further, it is well-known that equations (1) to (4) constitute a hyperbolic set of quasilinear first order partial differential equations [2]; hence there is a possibility that discontinuities in the dependent variables themselves develop in the solution. In the present paper, arguments are given for simple jump discontinuities in the first order derivatives of the flow variables, and is shown in § 4 that the discontinuities in the flow variable gradients satisfy a Riccati type equation [eq. (18)], which is amenable to analysis when transformed along the bicharacteristic direction.

3. Velocity of propagation

Let us consider a moving singularity surface $\Sigma(t)$ across which the flow variables are continuous but which is such that at least some of the first partial derivatives of these flow variables suffer jump discontinuities at the surface. Such a singularity surface is called a weak wave or a surface of weak discontinuity. We denote the jump in any flow quantity Z across $\Sigma(t)$ by $[Z] = Z_2 - Z_1$, where Z_1 is the value of Z immediately ahead of $\Sigma(t)$ and Z_2 is the value of Z immediately behind it. Suppose that $\Sigma(t)$ is given by $f(x_i, t) = 0$ and that we denote by $n_i = f_{,i}/|\operatorname{grad} f|$ the components of the unit normal vector and by $G = -(\partial f/\partial t)/|\operatorname{grad} f|$ the normal speed of propagation of $\Sigma(t)$. Then a result of [9] implies that

$$[Z_{,i}] = A n_i, [\partial Z/\partial t] = -AG, \tag{5, 6}$$

where Z stands for any of the flow variables, p, ρ, q, u_i and S and $A = [Z_{,i}] n_i$ is a scalar function defined on $\Sigma(t)$.

Taking jumps, across $\Sigma(t)$, in equations (1) to (4) and making use of (5) and (6), we get

$$U\xi = \rho_1 c_{f_1}^2 \lambda_i n_i, \rho_1 U\lambda_i = \xi n_i, \tag{7, 8}$$

$$U\chi = 0, \qquad U\theta = 0, \tag{9, 10}$$

where $U = (G - u_{n1})$ is the relative speed of advance of $\Sigma(t)$ into the gas and $\lambda_i = [u_{i,j}] n_j$, $\xi = [p_{,i}] n_i$, $\chi = [S_{,i}] n_i$, $\theta = [q_{,i}] n_i$ and $u_{n_1} = u_{i_1} n_i$ are the quantities defined on $\Sigma(t)$. Equations (7) to (10) yield the following relations:

$$\lambda_i = \lambda n_i, \qquad \lambda = \xi/\rho_1 U, \tag{11, 12}$$

$$\chi = 0, \qquad \theta = 0. \tag{13, 14}$$

In view of the relations (11, 12), equation (7) yields

$$G - u_{n_1} = \pm c_{f_1}.$$

The case $G = 0$ in which the surface moves with the fluid is discarded as having no physical interest. For an advancing wave surface, we shall take G to be positive and thus, we take

$$G - u_{n_1} = c_{f_1}, \tag{15}$$

Remark. Here we have used the theory of singular surfaces (see [9]) which in comparison to the theory of characteristics, quickly leads to the results of general significance. Of course, in a one-dimensional situation, the method of characteristics immediately leads to equations (12) to (15) as is shown by Chu ([2], p. 40), but in a 3-dimensional situation, which is under consideration, the corresponding derivation by using the method of characteristics would not be so simple as it seems in a one-dimensional case. In fact, if one uses the method of characteristics in a 3-dimensional situation, the idea of a characteristic curve is to be replaced by that of a characteristic surface along which an interior derivative involves derivatives in several directions in the surface unless the surface is plane or cylindrical (one-dimensional case) for which the differentiation is only in one direction viz., along the characteristic curve; this is the reason why 3-dimensional characteristic calculations are much more difficult than the one-dimensional case. Using the characteristics method, one can see after performing several manipulations that (15) is a root of a characteristic equation (see [6], pp. 33–38), and the eigenvector corresponding to this root yields the above equations (11) to (14). To make the analysis simple, and to arrive at the final results quickly we prefer here to make use of the theory of singular surfaces, as against the theory of characteristics.

4. Behaviour at the wave front

If we differentiate (1) and (2) partially with respect to x_k, take jumps across $\Sigma(t)$ and multiply the resulting equations by n_k, then, using the relations (6), (11), (13), (14), (15) and the second order compatibility conditions due to Thomas [9], we obtain

$$\frac{\delta\xi}{\delta t} + u^\alpha \xi_{,\alpha} - c_{f_1}(\bar{\xi} + c_{f_1}\zeta\lambda - \rho_1 c_{f_1}\bar{\lambda}) - 2\rho_1 c_{f_1}^2 \Omega\lambda - \Gamma_1 \lambda\xi$$
$$+ 2\Lambda_1\xi + \Gamma_1\xi (\partial u_i/\partial n)_1 n_i + \lambda\{(\partial p/\partial n)_1 + c_{f_1}^2 (\partial\rho/\partial n)_1$$
$$+ \rho_1 (\partial c_f^2/\partial n)_1\} + c_{f_1}^2 \zeta (\partial u_i/\partial n)_1 n_i = 0, \tag{16}$$

$$\rho_1 \frac{\delta\lambda}{\delta t} + \rho_1 u^\alpha \lambda_{,\alpha} + (\bar{\xi} + c_{f_1}\zeta\lambda - \rho_1 c_{f_1}\bar{\lambda}) - \rho_1\lambda^2 - c_{f_1}\lambda (\partial\rho/\partial n)_1$$
$$+ 2\rho_1\lambda (\partial u_i/\partial n)_1 n_i + \left(\frac{\partial u_i}{\partial t} + u_j u_{i,j}\right)_1 n_i\zeta = 0, \tag{17}$$

where $\Gamma_1 = 1 + \rho_1 (\partial c_f^2/\partial p)_1$, $\Lambda_1 = \frac{1}{2}\{\omega_1\sigma_1 (\partial c_f^2/\partial p)_1 + \omega_1 c_{f_1}^2 (\partial\sigma/\partial p)_1$
$+ \sigma_1 c_{f_1}^2 (\partial\omega/\partial p)_1\}$,

$\zeta = [p_{,i}] n_i$, $\bar{\lambda} = [u_{i,jk}] n_i n_j n_k$, $\bar{\xi} = [p_{,ij}] n_i n_j$,

$u^\alpha = u_i x_{i,\alpha}$, $2\Omega = g^{\alpha\beta} b_{\alpha\beta}$ and the subscript $(\ )_1$ indicates evaluation ahead of the wave surface $\Sigma(t)$. The quantity Ω defined above is the mean curvature of $\Sigma(t)$ with $g^{\alpha\beta}$ and $b_{\alpha\beta}$ being the first and second fundamental forms of $\Sigma(t)$. The quantities Γ_1 and Λ_1 defined above are functions of p_1, S_1 and q_1. For an

ideal gas, it can be readily verified that $\Gamma_1 = \gamma$, the adiabatic index. Eliminating λ and ξ from (16) and (17), we find, on using (12) and the fact that $c_{f_1}\zeta = \rho_1\lambda$, that

$$\rho_1 c_{f_1}\left(\frac{\delta\lambda}{\delta t} + u^\alpha \lambda_{,\alpha}\right) + \left(\frac{\delta\xi}{\delta t} + u^\alpha \xi_{,\alpha}\right) + P\zeta + Q\zeta^2 = 0\,, \tag{18}$$

where $$P = 2\Lambda_1 c_{f_1}^2 - 2c_{f_1}^3\Omega + 3c_{f_1}^2 (\partial u_i/\partial n)_1 n_i - \frac{c_{f_1}^3}{\rho_1}(\partial\rho/\partial n)_1$$

$$+ \frac{c_{f_1}}{\rho_1}\{(\partial p/\partial n)_1 + c_{f_1}^2 (\partial\rho/\partial n)_1 + \rho_1 (\partial c_{f_1}^2/\partial n)_1\} + \Gamma_1 c_{f_1}^2 (\partial u_i/\partial n)_1 n_i$$

$$+ c_{f_1}\left(\frac{\partial u_i}{\partial t} + u_j u_{i,j}\right)_1 n_i\,,$$

and $$Q = c_{f_1}^3 (1 + \Gamma_1)/\rho_1\,.$$

In view of relation (12) and $c_{f_1}\zeta = \rho_1\lambda$, equation (18) yields a differential equation for ζ, and, therefore, one for λ and one for ξ along orthogonal trajectories of $\Sigma(t)$. In equation (18) terms involving surface derivative cause some difficulty in its interpretation, but if we transform (18) to a differential equation along bicharacteristic curves, this difficulty disappears.

In order to proceed we need to recall that the bicharacteristic curve $x_i = x_i(t)$ in the characteristic manifold $\Sigma(t)$ for the system (1) to (4) satisfies equation [4]

$$\hat{d}\, x_i/dt = \frac{\partial F}{\partial\varphi_{,i}} \Big/ \frac{\partial F}{\partial\varphi_{,t}}\,, \tag{19}$$

where $\hat{d}/dt$ is the time derivative moving with the wave front along the bicharacteristic curve defined as

$$\frac{\hat{d}}{dt} = \frac{\partial}{\partial t} + v_i \frac{\partial}{\partial x_i}\,, \tag{20}$$

(v_i are the characteristic velocity components), a comma followed by the subscript t indicates partial differentiation with respect to time, and F is the characteristic determinant for the system which for the present case (i.e., $G - u_{n1} = c_{f_1}$) takes the form

$$F = \varphi_{,t} + u_i\varphi_{,i} + c_{f_1}(\varphi_{,i}\varphi_{,i})^{1/2}\,. \tag{21}$$

Equation (19) together with (21) yields

$$v_i = \hat{d}x_i/dt = u_i + c_{f_1} n_i\,. \tag{22}$$

Introducing the $\delta/\delta t$ derivative, it follows immediately from (22) that for any flow variable Z defined on $\Sigma(t)$, we have the relation

$$\hat{d}Z_i/dt = \frac{\delta Z}{\delta t} + u^\alpha Z_{,\alpha}\,. \tag{23}$$

Using (23) in (18) and substituting λ and ξ in terms of ζ from (12) and $c_{f_1}\zeta = \rho_1\lambda$, we get

$$\frac{\hat{d}\zeta}{dt} + \mu\zeta + \beta\zeta^2 = 0, \tag{24}$$

where $\mu = \frac{1}{2}\left\{\frac{\hat{d}}{dt}\log(c_{f_1}^3/\rho_1) + (P/c_{f_1}^2)\right\}$ and $\beta = \frac{1}{2}(Q/c_{f_1}^2)$.

Equation (24) is the transport equation governing the propagation of the discontinuity ζ, in fact, it dominates the behaviour of weak discontinuities of the solution of the quasi-linear system of hyperbolic equations (1) to (4). In the subsequent discussion we shall see how nicely this equation shows the interplay of damping and steepening (equivalently, blow-up) tendencies in the linear damping term $\mu\zeta$ and the nonlinear steepening term $\beta\zeta^2$. By the term 'steepening' or 'blow-up' of discontinuities we mean that because of the nonlinear term $\beta\zeta^2$, the discontinuities in the gradients of p, ρ and u_i become infinite in a finite time, and after this time a shock wave with discontinuities in the dependent variables themselves is introduced. In this context, we would like to refer to a recent and very interesting paper by John [7] who examines the problem of existence of global solutions of a nonlinear hyperbolic second order partial differential equation; as one might expect, the nonlinearity in this equation plays an important role in the existence of global solutions. In his analysis, the situation when the solution becomes unbounded at some finite time is identified with the non-existence of global solutions after that time or equivalently with the blow-up of solutions; he has analysed the situation under which a non-trivial solution with an initial data of compact support blows-up after a finite time. In the present paper, we have analysed in detail the solution of (24) with a discontinuous initial data at the boundary [which, in the present case, is the singularity surface $\Sigma(t)$], and we have established some explicit criteria for the decay or blow-up of discontinuities. Note that here, in contrast to John's paper, there is no problem of the existence of a solution of (24); the solution does exist but we are interested in studying the behaviour of the solution.

Equation (24) can be integrated to yield

$$\zeta = \frac{\zeta_0 (A/A_0)^{1/2} \exp[-\phi(t)]}{1 + \zeta_0 A_0^{1/2} I(t)}, \tag{25}$$

where $A = (c_{f_1}^3/\rho_1)$, $\phi(t) = \frac{1}{2}\int_0^t (P/c_{f_1}^2)\,dt$,

$$I(t) = \int_0^t (\beta/A)\exp[-\phi(\hat{t})]\,d\hat{t},$$

and the subscript zero indicates an initial value at $t = 0$. Equation (25) shows that if both $\phi(t)$ and $I(t)$ are continuous for $0 \leqslant t < \bar{t}$ and have finite limits $\phi(\bar{t})$, $I(\bar{t})$ as $t \to \bar{t}$ and if sign ζ_0 = sign $I(t)$ then the right hand side of (25) will not only remain continuous throughout $0 \leqslant t < \bar{t}$ but will also approach a finite limit as $t \to \bar{t}$. Also if sign $\zeta_0 = -$ sign $I(t)$, (25) will remain finite throughout

$[0, \bar{t}]$ provided $|\zeta_0| < \zeta_c^*$, where ζ_c^* is a positive quantity defined as $\zeta_c^* = (\Lambda_0^{1/2} |I(\infty)|)^{-1}$. But if sign $\zeta_0 = -$ sign $I(t)$ and $|\zeta_0| > \zeta_c^*$, then it follows from (25) that there will exist a time $t^* < \bar{t}$, given by $I(t^*) = -1/\zeta_0 \Lambda_0^{1/2}$, such that $\zeta \to \infty$ as $t \to t^*$. This signifies the appearance of a shock wave at an instant t^*. Further, if $|\zeta_0| = \zeta_c^*$ and sign $\zeta_0 = -$ sign $I(t)$, then we find that ζ is continuous for t in $[0, \bar{t})$ but approaches infinity as $t \to \bar{t}$.

5. Waves entering in a uniform region at rest

We shall now consider the case where since time zero the wave has been propagating into a region in which the gas is at rest. Then we find that for such a wave the differential equation (18) becomes

$$\frac{\delta \zeta}{\delta t} + (\Lambda_1 - c_{f_1} \Omega)\, \zeta + \frac{c_{f_1}(1 + \Gamma_1)}{2\rho_1} \zeta^2 = 0. \tag{26}$$

It is clear that the behaviour of the solutions of (26) will depend critically on the sign of Λ_1. It has been shown in [1] that for a state of strong equilibrium Λ_1 is non-negative whereas in a weak equilibrium state Λ_1 may be positive or negative.

The mean curvature Ω at any point of the wave surface $\Sigma(t)$ is given by [8]

$$\Omega = \frac{\Omega_0 - K_0 c_{f_1} t}{1 - 2\Omega_0 c_{f_1} t + K_0 c_{f_1}^2 t^2}, \tag{27}$$

where $\Omega_0 = \frac{1}{2}(k_1 + k_2)$ is the mean curvature and $K_0 = k_1 k_2$ is the Gaussian curvature of $\Sigma(t)$ at $t = 0$ with k_1 and k_2 being the principal curvatures. When k_1 and k_2 are both non-positive, the wave is divergent. On the other hand if one or both the principal curvatures are positive, then the wave is convergent. In the following section we consider these two cases under different thermodynamical situations namely the weak and strong equilibrium states.

Equation (26), after substitution from (27), can be integrated to yield

$$\zeta = \frac{\zeta_0 \exp(-\Lambda_1 t)\, I_1(t)}{1 + \dfrac{\zeta_0 (\Gamma_1 + 1) c_{f_1}}{2\rho_1} I_2(t)}, \tag{28}$$

where $$I_1(t) = \{(1 - k_1 c_{f_1} t)(1 - k_2 c_{f_1} t)\}^{-1/2}, \tag{29}$$

and $$I_2(t) = \int_0^t I_1(\hat{t}) \exp(-\Lambda_1 \hat{t})\, d\hat{t}. \tag{30}$$

5.1. *Diverging waves*

For diverging waves, both k_1 and k_2 are non-positive, and it is apparent from (29) and (30) that I_1 and I_2 both converge to finite limits as $t \to \infty$. Hence, it follows from (28) that, for $\Lambda_1 > 0$, an expansive wave front (i.e., $\zeta_0 > 0$) decays and damps out ultimately. But if $\zeta_0 < 0$ (i.e., a compressive wave front) then it

follows from (28) that there exists a positive critical value of the discontinuity ζ_c given by

$$\zeta_c = \left\{ \frac{(\Gamma_1 + 1)\, c_{f_1}}{2\rho_1} \int_0^\infty I_1(\hat{t}) \exp(-\Lambda_1 \hat{t})\, d\hat{t} \right\}^{-1}, \tag{31}$$

such that waves with initial discontinuity $|\zeta_0| < \zeta_c$ damp to zero and the waves with initial discontinuity $|\zeta_0| > \zeta_c$ grow without bound in a finite time t_c given by

$$I_2(t_c) = 2\rho_1 / c_{f_1} (\Gamma_1 + 1)\, |\zeta_0|. \tag{32}$$

Further, if $\zeta_0 < 0$ and $|\zeta_0| = \zeta_c$, then on using l' Hospital and Leibnitz rules, it follows from (28) that

$$|\zeta| \to 2\rho_1 \Lambda_1 / c_{f_1} (\Gamma_1 + 1)$$

as $t \to \infty$. Thus, a diverging compressive wave for $|\zeta_0| = \zeta_c$ can neither terminate into a shock wave nor can it ever completely damp out. From (31), it follows that $\partial \xi_c / \partial \Lambda_1 > 0$, which means that the critical value of the initial discontinuity increases with Λ_1. Also $\partial \zeta_c / |\partial k_1| > 0$ and $\partial \zeta_c / |\partial k_2| > 0$, which imply that the initial curvatures have a stabilising effect on the tendency of the wave surface to grow into a shock in the sense that an increase in the value of initial curvature causes an increase in the critical value. Relations (31) and (32) can be specialised to waves with plane ($k_1 = k_2 = 0$), cylindrical ($k_1 = -1/R_0$, $k_2 = 0$). and spherical ($k_1 = k_2 = -1/R_0$) geometry, where R_0 is the radius of the wave front at $t = 0$. We find that in these cases, the critical value of the initial discontinuity and the time taken for the shock formation are given by the following relations

Plane wave
$$\begin{cases} \zeta_c = 2\rho_1 \Lambda_1 / c_{f_1} (\Gamma_1 + 1) \\ t_c = \Lambda_1^{-1} \log \{1 - (2\rho_1 \Lambda_1 / c_{f_1} (\Gamma_1 + 1)\, |\zeta_0|)\}^{-1} \end{cases}$$

Cylindrical wave
$$\begin{cases} \zeta_c = \dfrac{2\rho_1}{c_{f_1}(\Gamma_1 + 1)} \left(\dfrac{\pi R_0}{\Lambda_1 c_{f_1}} \right)^{1/2} \dfrac{\exp(-\Lambda_1 R_0 / c_{f_1})}{\operatorname{erfc}(\Lambda_1 R_0 / c_{f_1})^{1/2}} \\ \operatorname{erfc}\left(\Lambda_1 t_c + \dfrac{\Lambda_1 R_0}{c_{f_1}} \right)^{1/2} = \left(1 - \dfrac{\zeta_c}{|\zeta_0|} \right) \operatorname{erfc}(\Lambda_1 R_0 / c_{f_1})^{1/2}, \end{cases}$$

Spherical wave
$$\begin{cases} \zeta_c = \dfrac{2\rho_1 \exp(-\Lambda_1 R_0 / c_{f_1})}{(\Gamma_1 + 1) R_0 E_i(\Lambda_1 R_0 / c_{f_1})} \\ E_i\left(\Lambda_1 t_c + \dfrac{\Lambda_1 R_0}{c_{f_1}} \right) = \left(1 - \dfrac{\zeta_c}{|\zeta_0|} \right) E_i(\Lambda_1 R_0 / c_{f_1}) \end{cases}$$

where $\operatorname{erfc}(x) = (2/\sqrt{\pi}) \int_x^\infty \exp(-t^2)\, dt$, $E_i(x) = \int_x^\infty t^{-1} \exp(-t)\, dt$

are, respectively, the complementary error function and the exponential integral.

For $A_1 < 0$ which corresponds to the case of a weak equilibrium state, it follows from (28) that for $\zeta_0 > 0$,

$$\zeta \to 2\rho_1 A_1 / c_{f_1}(\Gamma_1 + 1) \text{ as } t \to \infty,$$

i.e., all diverging expansive waves approach a limiting value. But if $\zeta_0 < 0$, then it follows from (28) that $\zeta \to \infty$ as $t \to \bar{t}_c$, where $\bar{t}_c$ is a finite time given by

$$\int_0^{\bar{t}_c} \exp(|A_1| t) I_1(t)\, dt = 2\rho_1 / c_{f_1}(\Gamma_1 + 1) |\zeta_0|.$$

Thus, in a weak equilibrium state all diverging compressive waves, no matter how weak initially, always end up into a shock after a finite time.

2. *Converging waves*

When only one of the initial principal curvatures is positive or both are positive and $k_1 \neq k_2$, in that case, for $A_1 > 0$ or $A_1 < 0$, there exists a finite time t^* given by the smallest positive root of $(1 - k_1 c_{f_1} t)(1 - k_2 c_{f_1} t) = 0$ such that $I_1(t) \to \infty$ as $t \to t^*$ whereas $I_2(t)$ tends to a finite value as $t \to t^*$. The fact that $I_2(t^*)$ is bounded follows from the argument that the singularity at t^* of the integrand in $I_2(t)$ is of the form $z^{-1/2} g(z)$ as $z \to 0$, where $g(z)$ is bounded. This can be easily seen by a suitable transformation $z = t - t^*$. Hence, it follows from (28) that for $\zeta_0 > 0$, $|\zeta| \to \infty$ as $t \to t^*$, i.e., a converging expansive wave forms a focus. But if $\zeta_0 < 0$, then it follows from (28) that there exists a critical value of the discontinuity $\hat{\zeta}_c$ given by

$$\hat{\zeta}_c = 2\rho_1 / c_{f_1}(\Gamma_1 + 1) I_2(t^*),$$

such that if $|\zeta_0| < \hat{\zeta}_c$, then $|\zeta| \to \infty$ as $t \to t^*$ which corresponds to the formation of a focus but not the shock. On the other hand if $|\zeta_0| > \hat{\zeta}_c$, then we find that $|\zeta| \to \infty$ as $t \to \hat{t}_c$, where $\hat{t}_c$ is finite time given by

$$I_2(\hat{t}_c) = 2\rho_1 / c_{f_1}(\Gamma_1 + 1) |\zeta_0|.$$

It is evident that $\hat{t}_c < t^*$. Thus, when $|\zeta_0| > \hat{\zeta}_c$, we find that the shock wave is formed before the focus can. The case $|\zeta_0| = \hat{\zeta}_c$ corresponds to the simultaneous formation of a shock and a focus.

(ii) When both the initial principal curvatures are positive and $k_1 = k_2$ in that case the singularity at t^* of the integrand in $I_2(t)$ is of the form $z^{-1} r(z)$ as $z \to 0$ where $r(z)$ is bounded away from zero. Hence, $I_2(t) \to \infty$ as $t \to t^*$. Thus it follows from (28) that if $\zeta_0 > 0$ then $|\zeta| \to \infty$ as $t \to t^*$, i.e., a focus is formed within a finite time t^*. But if $\zeta_0 < 0$, then there exists a finite time $\tilde{t} < t^*$ given

$$I_2(\tilde{t}) = 2\rho_1 / c_{f_1}(\Gamma_1 + 1) |\zeta_0|,$$

for which the denominator of (28) vanishes whereas the numerator remains finite. This means that, in this particular situation, a converging compressive wave,

For $A_1 < 0$ which corresponds to the case of a weak equilibrium state, it follows from (28) that for $\zeta_0 > 0$,

$$\zeta \to 2\rho_1 A_1 / c_{f_1} (\Gamma_1 + 1) \text{ as } t \to \infty,$$

i.e., all diverging expansive waves approach a limiting value. But if $\zeta_0 < 0$, then it follows from (28) that $\zeta \to \infty$ as $t \to \bar{t}_c$, where $\bar{t}_c$ is a finite time given by

$$\int_0^{\bar{t}_c} \exp(|A_1| t) I_1(t)\, dt = 2\rho_1 / c_{f_1} (\Gamma_1 + 1) |\zeta_0|.$$

Thus, in a weak equilibrium state all diverging compressive waves, no matter how weak initially, always end up into a shock after a finite time.

5.2. *Converging waves*

(i) When only one of the initial principal curvatures is positive or both are positive and $k_1 \neq k_2$, in that case, for $A_1 > 0$ or $A_1 < 0$, there exists a finite time t^* given by the smallest positive root of $(1 - k_1 c_{f_1} t)(1 - k_2 c_{f_1} t) = 0$ such that $I_1(t) \to \infty$ as $t \to t^*$ whereas $I_2(t)$ tends to a finite value as $t \to t^*$. The fact that $I_2(t^*)$ is bounded follows from the argument that the singularity at t^* of the integrand in $I_2(t)$ is of the form $z^{-1/2} g(z)$ as $z \to 0$, where $g(z)$ is bounded. This can be easily seen by a suitable transformation $z = t - t^*$. Hence, it follows from (28) that for $\zeta_0 > 0$, $|\zeta| \to \infty$ as $t \to t^*$, i.e., a converging expansive wave forms a focus. But if $\zeta_0 < 0$, then it follows from (28) that there exists a critical value of the discontinuity $\hat{\zeta}_c$ given by

$$\hat{\zeta}_c = 2\rho_1 / c_{f_1} (\Gamma_1 + 1) I_2(t^*),$$

such that if $|\zeta_0| < \hat{\zeta}_c$, then $|\zeta| \to \infty$ as $t \to t^*$ which corresponds to the formation of a focus but not the shock. On the other hand if $|\zeta_0| > \hat{\zeta}_c$, then we find that $|\zeta| \to \infty$ as $t \to \hat{t}_c$, where $\hat{t}_c$ is finite time given by

$$I_2(\hat{t}_c) = 2\rho_1 / c_{f_1} (\Gamma_1 + 1) |\zeta_0|.$$

It is evident that $\hat{t}_c < t^*$. Thus, when $|\zeta_0| > \hat{\zeta}_c$, we find that the shock wave is formed before the focus can. The case $|\zeta_0| = \hat{\zeta}_c$ corresponds to the simultaneous formation of a shock and a focus.

(ii) When both the initial principal curvatures are positive and $k_1 = k_2$ in that case the singularity at t^* of the integrand in $I_2(t)$ is of the form $z^{-1} r(z)$ as $z \to 0$ where $r(z)$ is bounded away from zero. Hence, $I_2(t) \to \infty$ as $t \to t^*$. Thus it follows from (28) that if $\zeta_0 > 0$ then $|\zeta| \to \infty$ as $t \to t^*$, i.e., a focus is formed within a finite time t^*. But if $\zeta_0 < 0$, then there exists a finite time $\tilde{t} < t^*$ given by

$$I_2(\tilde{t}) = 2\rho_1 / c_{f_1} (\Gamma_1 + 1) |\zeta_0|,$$

for which the denominator of (28) vanishes whereas the numerator remains finite. This means that, in this particular situation, a converging compressive wave,

Vol. 89, January–December 1980

SUBJECT INDEX (Mathematical Sciences)

irrespective of the thermodynamical state, whether weak or strong, always grows into a shock wave before the formation of the focus.

Acknowledgement

The authors are thankful to the referee for his valuable comments on the earlier version of the paper.

References

[1] Bowen R M 1969 *Arch. Ration. Mech. Anal.* **33** 169
[2] Chu B T 1970 *Non-equilibrium flows* (ed.) P P Wegner (New York : Marcel Dekker)
[3] Clarke J F and McChesney M 1976 *Dynamics of relaxing gases* (London : Butterworths)
[4] Courant R and Hilbert D 1962 *Methods of mathematical physics* (New York : Interscience)
[5] Elcrat A R 1977 *Int. J. Eng. Sci.* **15** 29
[6] Jeffrey A and Taniuti T 1964 *Nonlinear wave propagation* (New York : Academic Press)
[7] John F 1979 *Manuscripta Math.* **28** 235
[8] Thomas T Y 1965 *Concepts from tensor analysis and differential geometry* (New York : Academic Press)
[9] Thomas T Y 1966 *Int. J. Eng. Sci.* **4** 207

Vol. 89, January–December 1980

SUBJECT INDEX (Mathematical Sciences)

AUTHOR INDEX (Mathematical Sciences)